Oracle Datenbank

Enterprise applications

NetApp
February 10, 2026

Inhalt

Oracle Datenbank

Oracle-Datenbanken auf ONTAP

ONTAP wurde für Oracle Datenbanken entwickelt. Seit Jahrzehnten ist ONTAP für die speziellen Anforderungen relationaler Datenbank-I/O optimiert. Es wurden mehrere ONTAP-Funktionen speziell dafür entwickelt, die Anforderungen von Oracle Datenbanken zu bedienen – und sogar auf Wunsch von Oracle Inc. Selbst.

Diese Dokumentation ersetzt die zuvor veröffentlichten technischen Berichte *TR-3633: Oracle Databases on ONTAP; TR-4591: Oracle Data Protection: Backup, Recovery, Replizierung; TR-4592: Oracle on MetroCluster; und TR-4534: Migration von Oracle Databases to NetApp Storage Systems*

Neben den vielfältigen Möglichkeiten, die ONTAP für eine Datenbankumgebung bietet, gibt es auch zahlreiche Benutzeranforderungen, darunter Datenbankgröße, Performance-Anforderungen und Datensicherung. Bekannte Implementierungen von NetApp Storage umfassen alles von einer virtualisierten Umgebung mit ca. 6,000 Datenbanken unter VMware ESX bis hin zu einem Data Warehouse mit einer einzigen Instanz, das derzeit eine Größe von 996 TB aufweist und weiter wächst. Aus diesem Grund gibt es nur wenige klare Best Practices für die Konfiguration einer Oracle Datenbank auf NetApp Storage.

Die Anforderungen für den Betrieb einer Oracle Database auf NetApp Storage werden auf zweierlei Weise erfüllt. Erstens, wenn eine klare Best Practice besteht, wird sie ausdrücklich genannt. Im Allgemeinen werden viele Designüberlegungen erläutert, die von den Architekten von Oracle-Speicherlösungen auf der Grundlage ihrer spezifischen Geschäftsanforderungen berücksichtigt werden müssen.

ONTAP Konfiguration auf AFF/ FAS -Systemen

RAID

RAID bezieht sich auf den Einsatz von Redundanz, um Daten vor dem Verlust eines Laufwerks zu schützen.

Gelegentlich stellen sich Fragen zu RAID-Levels bei der Konfiguration von NetApp-Speicher, der für Oracle-Datenbanken und andere Enterprise-Applikationen verwendet wird. Viele, von Oracle bewährte Verfahren zur Storage Array-Konfiguration enthalten Warnungen über die Verwendung von RAID-Spiegelung und/oder Vermeidung bestimmter Arten von RAID. Obwohl in ihnen gültige Punkte aufgeführt sind, gelten diese Quellen nicht für RAID 4 und die in ONTAP verwendeten NetApp RAID DP und RAID-TEC Technologien.

RAID 4, RAID 5, RAID 6, RAID DP und RAID-TEC nutzen Parität, um sicherzustellen, dass es bei einem Laufwerksausfall zu keinem Datenverlust kommt. Diese RAID-Optionen bieten im Vergleich zur Spiegelung eine viel bessere Speichernutzung, aber die meisten RAID-Implementierungen haben einen Nachteil, der Schreibvorgänge beeinträchtigt. Für den Abschluss eines Schreibvorgangs in anderen RAID-Implementierungen sind möglicherweise mehrere Laufwerkszugriffe erforderlich, um die Paritätsdaten neu zu generieren. Dies ist ein Prozess, der allgemein als RAID-Abzug bezeichnet wird.

Bei ONTAP fallen jedoch keine RAID-Einbußen an. Dies liegt an der Integration von NetApp WAFL (Write Anywhere File Layout) mit der RAID-Schicht. Schreibvorgänge werden im RAM zusammengeführt und als vollständiger RAID-Stripe einschließlich der Paritätsgenerierung vorbereitet. ONTAP muss für einen Schreibvorgang keinen Lesevorgang durchführen. Das bedeutet, dass ONTAP und WAFL die RAID-Einbußen

vermeiden. Die Performance für latenzkritische Vorgänge, wie die Protokollierung von Wiederherstellungen, wird ohne Behinderung durchgeführt und zufällige Schreibvorgänge von Datendateien verursachen keine RAID-Beeinträchtigungen, da die Parität neu generiert werden muss.

In Bezug auf statistische Zuverlässigkeit bietet selbst RAID DP besseren Schutz als RAID Mirroring. Das Hauptproblem besteht in der Nachfrage nach Laufwerken während einer RAID-Wiederherstellung. Bei einem gespiegelten RAID-Satz ist das Risiko eines Datenverlusts aufgrund eines Laufwerksausfalls bei der Wiederherstellung an seinen Partner im RAID-Satz deutlich größer als das Risiko eines dreifachen Laufwerksausfalls in einem RAID DP-Satz.

Kapazitätsmanagement

Für das Management von Datenbanken oder anderen Enterprise-Applikationen mit vorhersehbarem, leicht verwaltbarem, hochperformantem Enterprise-Storage ist auf den Laufwerken freier Speicherplatz für das Daten- und Metadaten-Management erforderlich. Die Menge des freien Speicherplatzes hängt vom Typ des verwendeten Laufwerks und von den Geschäftsprozessen ab.

Der freie Speicherplatz wird definiert als jeder Speicherplatz, der nicht für tatsächliche Daten verwendet wird. Er umfasst nicht zugewiesenen Speicherplatz im Aggregat selbst und ungenutzten Speicherplatz innerhalb der einzelnen Volumes. Thin Provisioning ist ebenfalls zu berücksichtigen. Ein Volume kann beispielsweise eine LUN mit 1 TB enthalten, von der nur 50 % von echten Daten genutzt werden. In einer Thin Provisioning Umgebung wird hierdurch scheinbar 500 GB Speicherplatz belegt. In einer vollständig bereitgestellten Umgebung scheint jedoch die volle Kapazität von 1 TB genutzt zu sein. Der nicht zugewiesene Speicherplatz von 500 GB ist ausgeblendet. Dieser Platz wird von tatsächlichen Daten nicht genutzt und sollte daher bei der Berechnung des gesamten freien Speicherplatzes berücksichtigt werden.

NetApp Empfehlungen für Storage-Systeme, die für Enterprise-Applikationen verwendet werden, sind wie folgt:

SSD-Aggregate, einschließlich AFF Systeme

NetApp empfiehlt mindestens 10% freien Platz. Dazu gehört der gesamte ungenutzte Speicherplatz, einschließlich freiem Speicherplatz innerhalb des Aggregats oder eines Volumes und sämtlicher freier Speicherplatz, der aufgrund der vollständigen Bereitstellung zugewiesen wird, aber nicht von tatsächlichen Daten genutzt wird. Logischer Speicherplatz ist dabei unwichtig. Die Frage lautet, wie viel tatsächlich freier physischer Speicherplatz für Daten zur Verfügung steht.

Die Empfehlung von 10 % freiem Platz ist sehr konservativ. SSD-Aggregate können Workloads mit noch höherer Auslastung ohne Auswirkungen auf die Performance unterstützen. Wenn die Auslastung des Aggregats jedoch nicht sorgfältig überwacht wird, steigt auch das Risiko, dass der Speicherplatz nicht ausgelastet wird. Darüber hinaus kann es bei der Ausführung eines Systems mit einer Kapazität von 99 % nicht zu einer Performance-Beeinträchtigung kommen, doch wäre damit wahrscheinlich ein Management-Aufwand verbunden, um zu verhindern, dass das System während der Bestellung zusätzlicher Hardware vollständig gefüllt wird. Zudem kann es einige Zeit dauern, bis zusätzliche Laufwerke beschaffen und installiert sind.

HDD-Aggregate, einschließlich Flash Pool Aggregaten

NetApp empfiehlt mindestens 15% freien Speicherplatz, wenn rotierende Laufwerke verwendet werden. Dazu gehört der gesamte ungenutzte Speicherplatz, einschließlich freiem Speicherplatz innerhalb des Aggregats oder eines Volumes und sämtlicher freier Speicherplatz, der aufgrund der vollständigen Bereitstellung zugewiesen wird, aber nicht von tatsächlichen Daten genutzt wird. Die Performance wird beeinträchtigt, da der freie Speicherplatz sich etwa bei 10 % befindet.

Storage Virtual Machines

Das Storage-Management für Oracle Datenbanken wird auf einer Storage Virtual Machine (SVM) zentralisiert.

Eine SVM, in der ONTAP CLI als vServer bezeichnet, ist eine grundlegende Funktionseinheit des Storage. Es ist hilfreich, eine SVM mit einem Gast auf einem VMware ESX Server zu vergleichen.

Bei der Erstinstallation verfügt ESX über keine vorkonfigurierten Funktionen, wie z. B. das Hosten eines Gastbetriebssystems oder die Unterstützung einer Endbenutzeranwendung. Es ist ein leerer Container, bis eine Virtual Machine (VM) definiert ist. ONTAP ist ähnlich. Die erste Installation von ONTAP umfasst keine Datenserverfunktionen, bis eine SVM erstellt wurde. Die Datenservices werden von der SVM-Persönlichkeit definiert.

Wie bei anderen Aspekten der Storage-Architektur hängen die besten Optionen für das Design von SVMs und Logical Interface (LIF) stark von den Skalierungsanforderungen und geschäftlichen Anforderungen ab.

SVMs

Es gibt keine offizielle Best Practice für die Bereitstellung von SVMs für ONTAP. Der richtige Ansatz hängt von Management- und Sicherheitsanforderungen ab.

Die meisten Kunden betreiben für die meisten ihrer täglichen Anforderungen eine primäre SVM, erstellen jedoch für besondere Anforderungen eine geringe Anzahl an SVMs. Sie können beispielsweise Folgendes erstellen:

- Eine SVM für eine kritische Geschäftsdatenbank, die von einem Expertenteam gemanagt wird
- Eine SVM für eine Entwicklungsgruppe, der eine vollständige administrative Kontrolle gegeben wurde, damit sie ihren eigenen Storage unabhängig managen können
- Eine SVM für sensible Geschäftsdaten wie Personaldaten oder Daten für Finanzberichte, für die das Administrationsteam begrenzt werden muss

In einer mandantenfähigen Umgebung können die Daten jedes Mandanten eine dedizierte SVM zugewiesen werden. Die Obergrenze für die Anzahl der SVMs und LIFs pro Cluster, HA-Paar und Node ist abhängig vom verwendeten Protokoll, dem Node-Modell und der Version von ONTAP. Konsultieren Sie die "NetApp Hardware Universe" Für diese Grenzwerte.

Performance-Management mit ONTAP QoS

Für die sichere und effiziente Verwaltung mehrerer Oracle Datenbanken ist eine effektive QoS-Strategie erforderlich. Der Grund dafür sind die stetig wachsenden Performance-Möglichkeiten eines modernen Storage-Systems.

Insbesondere die zunehmende Verbreitung von All-Flash-Storage ermöglicht die Konsolidierung von Workloads. Storage-Arrays mit rotierenden Medien unterstützten in der Regel nur eine begrenzte Anzahl I/O-

intensiver Workloads, da die IOPS-Funktionen einer älteren Rotationslaufwerkstechnologie begrenzt sind. Ein oder zwei hochaktive Datenbanken würden die zugrunde liegenden Laufwerke lange sättigen, bevor die Storage-Controller ihre Grenzen erreichten. Das hat sich geändert. Die Performance-Fähigkeit einer relativ kleinen Anzahl von SSD-Laufwerken kann sogar die leistungsstärksten Storage-Controller auslasten. So können Sie alle Funktionen der Controller nutzen, ohne die Gefahr eines plötzlichen Performance-Einbruchs aufgrund von Latenzspitzen der rotierenden Medien befürchten zu müssen.

Ein einfaches HA-AFF A800 System mit zwei Nodes kann bis zu eine Million zufällige IOPS verarbeiten, bevor die Latenz auf über eine Millisekunde steigt. Für sehr wenige einzelne Workloads wird erwartet, dass sie ein solches Niveau erreichen. Die vollständige Nutzung dieses AFF A800 System-Arrays beinhaltet das Hosting mehrerer Workloads. Für eine sichere Durchführung sind QoS-Kontrollen erforderlich, um die Vorhersehbarkeit zu gewährleisten.

ONTAP bietet zwei Arten von Quality of Service (QoS): IOPS und Bandbreite. QoS-Steuerungen können auf SVMs, Volumes, LUNs und Dateien angewendet werden.

IOPS QoS

Eine Steuerung der IOPS-QoS basiert offensichtlich auf dem IOPS-Wert einer bestimmten Ressource. Es gibt jedoch eine Reihe von Aspekten der IOPS-QoS, die möglicherweise nicht intuitiv sind. Einige Kunden waren anfangs verwirrt über den scheinbaren Anstieg der Latenz, wenn ein IOPS-Schwellenwert erreicht wurde. Die steigende Latenz ist das natürliche Ergebnis der IOPS-Begrenzung. Logischerweise funktioniert es ähnlich wie ein Token-System. Wenn beispielsweise ein bestimmtes Volume mit Datendateien über ein Limit von 10.000 IOPS verfügt, muss jede eintreffende I/O zuerst ein Token erhalten, um die Verarbeitung fortzusetzen. Solange in einer bestimmten Sekunde nicht mehr als 10.000 Token verbraucht wurden, sind keine Verzögerungen vorhanden. Wenn I/O-Vorgänge auf den Erhalt ihres Tokens warten müssen, wird diese Wartezeit als zusätzliche Latenz angezeigt. Je schwieriger eine Workload das QoS-Limit erreicht, desto länger muss jede I/O in der Warteschlange warten, bis sie verarbeitet wird. Dies scheint für den Benutzer eine höhere Latenz zu sein.

> Gehen Sie vorsichtig vor, wenn Sie QoS-Kontrollen auf Datenbanktransaktions-/Wiederherstellungsprotokolldaten anwenden. Die Performance-Anforderungen der Wiederherstellungsprotokollierung sind in der Regel viel, viel niedriger als Datendateien, jedoch erfolgt die Aktivität des Wiederherstellungsprotokolls sprunghafter. Die I/O-Vorgänge erfolgen in kurzen Impulsen, und ein QoS-Limit, das für die durchschnittlichen Wiederherstellungs-I/O-Level angemessen erscheint, kann für die tatsächlichen Anforderungen zu niedrig sein. Das Ergebnis können schwerwiegende Performance-Einschränkungen sein, wenn die QoS sich mit jedem Burst des Wiederherstellungsprotokolls befasst. Die Wiederherstellungs- und Archivprotokollierung sollte im Allgemeinen nicht durch QoS beschränkt sein.

Bandbreiten QoS

Nicht alle I/O-Größen sind gleich. Beispielsweise führt eine Datenbank möglicherweise eine große Anzahl kleiner Blocklesevorgänge durch, wodurch der IOPS-Schwellenwert erreicht wird. Datenbanken führen aber möglicherweise auch einen vollständigen Tabellenscan durch, der aus einer sehr kleinen Anzahl an großen Blocklesevorgängen bestehen würde, die eine sehr große Menge an Bandbreite verbrauchen, aber relativ wenig IOPS.

Ebenso könnte eine VMware Umgebung eine sehr hohe Anzahl zufälliger IOPS während des Startvorgangs verursachen, würde jedoch während eines externen Backups weniger, aber größere I/O-Vorgänge ausführen.

Manchmal erfordert ein effektives Performance-Management entweder Einschränkungen für die IOPS-Leistung oder die Bandbreite oder sogar beides.

Minimum/garantierte QoS

Viele Kunden wünschen sich eine Lösung mit garantierter QoS, was schwieriger zu erreichen ist, als sie möglicherweise verschwenderisch erscheint. Wenn Sie beispielsweise 10 Datenbanken mit einer Garantie von 10.000 IOPS platzieren, müssen Sie ein System für ein Szenario dimensionieren, in dem alle 10 Datenbanken gleichzeitig mit 10.000 IOPS, also insgesamt 100.000, laufen.

Eine minimale QoS-Steuerung soll am besten zum Schutz kritischer Workloads eingesetzt werden. Ein Beispiel wäre ein ONTAP Controller mit maximal 500.000 IOPS und einer Kombination aus Produktions- und Entwicklungs-Workloads. Sie sollten maximale QoS-Richtlinien auf Entwicklungs-Workloads anwenden, um zu verhindern, dass eine gegebene Datenbank den Controller für sich einmonopolisiert. Sie würden dann minimale QoS-Richtlinien auf Produktions-Workloads anwenden, um sicherzustellen, dass immer die erforderlichen IOPS bei Bedarf zur Verfügung stehen.

Anpassungsfähige QoS

Adaptive QoS bezeichnet die ONTAP-Funktion, bei der die QoS-Begrenzung auf der Kapazität des Storage-Objekts basiert. Sie wird selten bei Datenbanken eingesetzt, da zwischen der Größe einer Datenbank und ihren Performance-Anforderungen in der Regel keine Verknüpfung besteht. Große Datenbanken können nahezu inert sein, während kleinere Datenbanken die IOPS-intensivsten sein können.

Adaptive QoS kann bei Virtualisierungs-Datastores sehr hilfreich sein, da die IOPS-Anforderungen solcher Datensätze in der Regel mit der Gesamtgröße der Datenbank korrelieren. Ein neuerer Datenspeicher mit 1 TB VMDK-Dateien benötigt wahrscheinlich etwa die Hälfte der Performance als 2-TB-Datenspeicher. Durch anpassungsfähige QoS können Sie die QoS-Limits automatisch vergrößern, wenn der Datastore mit Daten gefüllt wird.

Effizienz

Die Funktionen zur Steigerung der Speicherplatzeffizienz von ONTAP sind für Oracle Datenbanken optimiert. In fast allen Fällen besteht der beste Ansatz darin, die Standardeinstellungen bei aktivierten Effizienzfunktionen zu belassen.

Funktionen für Platzeffizienz wie Komprimierung, Data-Compaction und Deduplizierung sind darauf ausgelegt, die Menge der logischen Daten zu einer bestimmten Menge des physischen Storage zu erhöhen. Das Ergebnis sind niedrigere Kosten und geringerer Management-Overhead.

Auf hohem Niveau ist Komprimierung ein mathematischer Prozess, bei dem Muster in Daten erkannt und so kodiert werden, dass der Platzbedarf reduziert wird. Dagegen erkennt die Deduplizierung tatsächlich wiederholte Datenblöcke und entfernt die fremden Kopien. Durch Data-Compaction können mehrere logische Datenblöcke denselben physischen Block auf den Medien gemeinsam nutzen.

In den nachfolgenden Abschnitten zu Thin Provisioning finden Sie eine Erläuterung des Wechselspiels zwischen Storage-Effizienz und fraktionaler Reservierung.

Komprimierung

Vor der Verfügbarkeit von All-Flash-Storage-Systemen war die Array-basierte Komprimierung nur eingeschränkt verfügbar, da die meisten I/O-intensiven Workloads eine sehr große Anzahl von Spindeln erforderten, um eine akzeptable Performance zu erreichen. Als Nebeneffekt der großen Anzahl von Laufwerken enthielten Storage-Systeme grundsätzlich viel mehr Kapazität als erforderlich. Mit dem Trend hin zu Solid-State-Storage hat sich die Situation verändert. Eine enorme Überprovisionierung von Laufwerken entfällt, nur weil eine gute Performance erzielt werden kann. Der Speicherplatz in einem Storage-System kann

den tatsächlichen Kapazitätsanforderungen angepasst werden.

Die gesteigerte IOPS-Fähigkeit von Solid-State-Laufwerken (SSDs) bringt im Vergleich zu rotierenden Laufwerken fast immer Kosteneinsparungen mit sich. Allerdings kann die Komprimierung durch eine höhere effektive Kapazität von Solid-State-Medien weitere Einsparungen erzielen.

Es gibt verschiedene Möglichkeiten, Daten zu komprimieren. Viele Datenbanken verfügen über eigene Komprimierungsfunktionen, dies wird jedoch in Kundenumgebungen selten beobachtet. Der Grund dafür ist in der Regel die Performance-Einbußen bei einem **Wechsel** zu komprimierten Daten. Bei einigen Anwendungen fallen zudem hohe Lizenzierungskosten für die Komprimierung auf Datenbankebene an. Und schließlich gibt es noch die allgemeinen Performance-Auswirkungen auf die Datenbankvorgänge. Es macht wenig Sinn, für eine CPU, die Datenkomprimierung und -Dekomprimierung durchführt, hohe Lizenzkosten pro CPU zu zahlen, anstatt eine echte Datenbankarbeit zu erledigen. Eine bessere Option ist, die Komprimierungsarbeiten auf das Storage-System zu verlagern.

Anpassungsfähige Komprimierung

Die adaptive Komprimierung wurde vollständig mit Enterprise-Workloads getestet, ohne dabei die Performance zu beeinträchtigen – selbst in einer All-Flash-Umgebung, in der die Latenz im Mikrosekunden-Bereich gemessen wird. Einige Kunden haben bei Verwendung der Komprimierung sogar eine Performance-Steigerung festgestellt, da die Daten im Cache komprimiert bleiben. Dadurch konnte die Menge des verfügbaren Cache in einem Controller erhöht werden.

ONTAP managt physische Blöcke in 4-KB-Einheiten. Die anpassungsfähige Komprimierung verwendet eine Standardkomprimierung von 8 KB. Dies bedeutet, dass Daten in 8-KB-Einheiten komprimiert werden. Dies entspricht der 8-KB-Blockgröße, die von relationalen Datenbanken am häufigsten verwendet wird. Kompressionsalgorithmen werden effizienter, da mehr Daten als eine Einheit komprimiert werden. Eine Komprimierungs-Blockgröße von 32 KB wäre speichereffizienter als eine Komprimierungsblockeinheit mit 8 KB. Das bedeutet, dass die adaptive Komprimierung bei Verwendung der standardmäßigen 8-KB-Blockgröße zu etwas niedrigeren Effizienzraten führt, jedoch bietet die Verwendung kleinerer Blockgrößen zur Komprimierung auch einen signifikanten Vorteil. Datenbank-Workloads umfassen einen großen Anteil an Überschreibungsaktivitäten. Beim Überschreiben eines komprimierten 32-KB-Datenblocks müssen die gesamten 32-KB-Daten zurückgelesen, dekomprimiert, der erforderliche 8-KB-Bereich aktualisiert, neu komprimiert und dann die gesamten 32-KB-Daten wieder auf die Laufwerke geschrieben werden. Dies ist für ein Storage-System ein sehr teurer Vorgang und der Grund dafür, dass bei einigen Storage Arrays anderer Anbieter, die auf größeren Komprimierungsblockgrößen basieren, auch die Performance bei Datenbank-Workloads erheblich beeinträchtigt wird.

Die von der anpassungsfähigen Komprimierung verwendete Blockgröße kann auf bis zu 32 KB gesteigert werden. Dies kann die Speichereffizienz verbessern und sollte bei stillgelegten Dateien wie Transaktionsprotokollen und Backup-Dateien in Betracht gezogen werden, wenn eine große Menge solcher Daten auf dem Array gespeichert wird. In manchen Situationen profitieren aktive Datenbanken mit 16-KB- oder 32-KB-Blockgröße möglicherweise auch von der Erhöhung der Blockgröße der anpassungsfähigen Komprimierung. Wenden Sie sich an einen Mitarbeiter von NetApp oder einen unserer Partner, um Rat zu erhalten, ob diese Lösung für Ihren Workload geeignet ist.

Blockgrößen der Komprimierung von mehr als 8 KB sollten nicht zusammen mit der Deduplizierung an Streaming-Backup-Zielen verwendet werden. Der Grund dafür ist, dass kleine Änderungen an den gesicherten Daten das 32-KB-Komprimierungsfenster beeinflussen. Wenn sich das Fenster verschiebt, unterscheiden sich die resultierenden komprimierten Daten in der gesamten Datei. Die Deduplizierung erfolgt nach der Komprimierung. Das heißt, die Deduplizierungs-Engine sieht jedes komprimierte Backup unterschiedlich. Wenn eine Deduplizierung von Streaming-Backups erforderlich ist, sollte nur eine blockadaptive Komprimierung von 8 KB verwendet werden. Die adaptive Komprimierung ist vorzuziehen, da sie bei kleineren Blöcken arbeitet und die Deduplizierungseffizienz nicht stört. Aus ähnlichen Gründen wirkt sich die Host-seitige Komprimierung auch in die Effizienz der Deduplizierung aus.

Kompressionsausrichtung

Die anpassungsfähige Komprimierung in einer Datenbankumgebung erfordert bestimmte Überlegungen zur Blockausrichtung der Komprimierung. Dies ist nur für Daten relevant, die Random Überschreibungen sehr spezifischer Blöcke unterliegen. Dieser Ansatz ähnelt im Konzept der gesamten Filesystem-Ausrichtung, wobei der Beginn eines Dateisystems an einer Grenze von 4K-Geräten ausgerichtet werden muss und die Blockgröße eines Dateisystems ein Vielfaches von 4K sein muss.

Ein Schreibvorgang von 8 KB in eine Datei wird beispielsweise nur komprimiert, wenn er an einer 8-KB-Grenze innerhalb des Dateisystems selbst ausgerichtet ist. Dieser Punkt bedeutet, dass er auf die ersten 8 KB der Datei, die zweiten 8 KB der Datei usw. fallen muss. Der einfachste Weg, um eine korrekte Ausrichtung zu gewährleisten, ist die Verwendung des korrekten LUN-Typs. Jede erstellte Partition sollte einen Offset vom Anfang des Geräts an haben, der ein Vielfaches von 8K ist, und eine Dateisystem-Blockgröße verwenden, die ein Vielfaches der Datenbank-Blockgröße ist.

Daten wie Backups oder Transaktions-Logs werden sequenziell geschrieben und umfassen mehrere Blöcke. Alle Blöcke werden komprimiert. Daher besteht keine Notwendigkeit, eine Ausrichtung zu erwägen. Das einzige I/O-Muster, das Bedenken aushinsichtlich des zufälligen Überschreibens von Dateien hat, ist das zufällige Überschreiben von Dateien.

Data-Compaction

Data-Compaction ist eine Technologie, die die Komprimierungseffizienz verbessert. Wie bereits erwähnt, erzielt die anpassungsfähige Komprimierung allein schon Einsparungen von 2:1, da sie auf das Speichern eines 8-KB-I/O-Blocks in einem 4-KB-WAFL-Block beschränkt ist. Komprimierungsmethoden mit größeren Blockgrößen verbessern die Effizienz. Sie sind jedoch nicht für Daten geeignet, die mit Überschreibungen kleiner Blöcke verbunden sind. Die Dekomprimierung von 32-KB-Dateneinheiten durch die Aktualisierung eines 8-KB-Abschnitts, die Datenkomprimierung und das Zurückschreiben auf die Laufwerke verursacht Overhead.

Data-Compaction sorgt dafür, dass mehrere logische Blöcke innerhalb physischer Blöcke gespeichert werden können. Beispielsweise kann eine Datenbank mit stark komprimierbaren Daten wie Text oder teilweise vollständigen Blöcken von 8 KB bis 1 KB komprimieren. Ohne Data-Compaction belegen diese 1 KB Daten immer noch einen gesamten 4-KB-Block. Durch die Inline-Data-Compaction können 1 KB komprimierte Daten zusammen mit anderen komprimierten Daten auf nur 1 KB physischen Speicherplatz gespeichert werden. Es handelt sich nicht um eine Komprimierungstechnologie. Es ist einfach eine effizientere Möglichkeit, Speicherplatz auf den Laufwerken zuzuweisen und sollte daher keine erkennbaren Performance-Auswirkungen verursachen.

Der Grad der erzielten Einsparungen variiert. Bereits komprimierte oder verschlüsselte Daten können in der Regel nicht weiter komprimiert werden. Daher profitieren diese Datensätze von der Data-Compaction nicht. Im Gegensatz dazu werden neu initialisierte Datendateien, die nur wenig mehr als Block-Metadaten und Nullen enthalten, mit bis zu 80 komprimiert.

Temperaturempfindliche Speichereffizienz

Temperaturempfindliche Speichereffizienz (TSSE) ist ab ONTAP 9.8 verfügbar. Es basiert auf Block-Zugriffs-Heatmaps, um selten genutzte Blöcke zu identifizieren und sie effizienter zu komprimieren.

Deduplizierung

Deduplizierung ist die Entfernung von Blockduplikaten aus einem Datensatz. Wenn beispielsweise derselbe 4-KB-Block in 10 verschiedenen Dateien vorhanden war, leitet die Deduplizierung diesen 4-KB-Block innerhalb aller 10 Dateien auf denselben physischen 4-KB-Block um. Im Ergebnis würde sich die Effizienz dieser Daten um 10:1 verbessern.

Daten wie Boot-LUNs von VMware lassen sich in der Regel sehr gut deduplizieren, da sie aus mehreren Kopien derselben Betriebssystemdateien bestehen. Es wurde eine Effizienz von 100:1 und höher festgestellt.

Einige Daten enthalten keine Datenduplikate. Ein Oracle-Block enthält beispielsweise einen Header, der global nur für die Datenbank gilt, und einen Trailer, der fast einzigartig ist. Aus diesem Grund führt die Deduplizierung einer Oracle Database selten zu Einsparungen von mehr als 1 %. Die Deduplizierung mit MS SQL Datenbanken ist etwas besser, aber eindeutige Metadaten auf Blockebene stellen immer noch eine Einschränkung dar.

In einigen Fällen wurde eine Speicherersparnis von bis zu 15 % bei Datenbanken mit 16 KB und großen Blockgrößen beobachtet. Die ersten 4-KB-Blöcke enthalten die global eindeutige Kopfzeile, und der letzte 4-KB-Block enthält den nahezu einzigartigen Trailer. Die internen Blöcke eignen sich für eine Deduplizierung, obwohl dies in der Praxis fast vollständig der Deduplizierung von gelöschten Daten zugeordnet ist.

Viele Arrays anderer Anbieter behaupten, Datenbanken unter der Annahme zu deduplizieren, dass eine Datenbank mehrfach kopiert wird. In dieser Hinsicht kann auch NetApp Deduplizierung eingesetzt werden, allerdings bietet ONTAP die bessere Option: NetApp FlexClone Technologie. Das Endergebnis ist das gleiche. Es werden mehrere Kopien einer Datenbank erstellt, die die meisten zugrunde liegenden physischen Blöcke nutzen. Ein Einsatz von FlexClone ist wesentlich effizienter, als Datenbankdateien zu kopieren und anschließend zu deduplizieren. Der Effekt ist die Nichtdeduplizierung und nicht die Deduplizierung, da ein Duplikat von vornirgends erstellt wird.

Effizienz und Thin Provisioning

Effizienzfunktionen sind Formen von Thin Provisioning. Beispielsweise kann eine 100-GB-LUN, die ein 100-GB-Volume belegt, bis zu 50 GB komprimiert werden. Es wurden noch keine tatsächlichen Einsparungen realisiert, da das Volume noch 100 GB beträgt. Das Volume muss zunächst verkleinert werden, damit der eingesparte Speicherplatz an anderer Stelle im System genutzt werden kann. Wenn spätere Änderungen an der 100GB-LUN dazu führen, dass die Daten weniger komprimierbar werden, dann vergrößert sich die LUN und das Volume könnte sich füllen.

Thin Provisioning wird nachdrücklich empfohlen, da es das Management vereinfachen und gleichzeitig eine deutliche Verbesserung der nutzbaren Kapazität mit den damit verbundenen Kosteneinsparungen ermöglichen kann. Der Grund hierfür ist einfach: Datenbankumgebungen enthalten oft viel leeren Speicherplatz, eine große Anzahl an Volumes und LUNs sowie komprimierbare Daten. Durch Thick Provisioning wird Speicherplatz auf Storage für Volumes und LUNs reserviert, für den Fall, dass sie eines Tages zu 100 % voll werden und 100 % nicht komprimierbare Daten enthalten. Das wird wohl nie passieren. Dank Thin Provisioning kann dieser Speicherplatz zurückgewonnen und an anderer Stelle verwendet werden. Das Kapazitätsmanagement kann auf dem Storage-System selbst basieren, anstatt auf vielen kleineren Volumes und LUNs.

Einige Kunden bevorzugen Thick Provisioning entweder für bestimmte Workloads oder generell basierend auf bestehenden Betriebs- und Beschaffungsmethoden.

Bei einem Volume mit Thick Provisioning müssen unbedingt alle Effizienzfunktionen des Volumes deaktiviert werden, einschließlich der Dekomprimierung und der Entfernung der Deduplizierung mit dem `sis undo` Befehl. Die Lautstärke sollte nicht in der Ausgabe angezeigt `volume efficiency show` werden. Ist dies der Fall, ist das Volume für Effizienzfunktionen noch teilweise konfiguriert. Daher funktionieren Überschreibungsgarantien anders. Dies erhöht die Wahrscheinlichkeit, dass Konfigurationsübersehungen dazu führen, dass das Volume unerwartet aus dem Speicherplatz kommt und zu Datenbank-I/O-Fehlern führt.

Best Practices für Effizienz

NetApp empfiehlt Folgendes:

AFF-Standards

Volumes, die auf ONTAP erstellt wurden und auf einem rein Flash-basierten AFF System ausgeführt werden, werden über Thin Provisioning mit allen Inline-Effizienzfunktionen bereitgestellt. Obwohl Datenbanken im Allgemeinen nicht von der Deduplizierung profitieren und nicht komprimierbare Daten enthalten können, sind die Standardeinstellungen dennoch für fast alle Workloads geeignet. ONTAP wurde mit dem Ziel entwickelt, alle Arten von Daten und I/O-Muster effizient zu verarbeiten. Dabei spielt es keine Rolle, ob es zu Einsparungen kommt oder nicht. Standardwerte sollten nur dann geändert werden, wenn die Gründe vollständig verstanden sind und es einen Vorteil gibt, dass sie abweichen.

Allgemeine Empfehlungen

- Wenn Volumes und/oder LUNs nicht über Thin Provisioning bereitgestellt werden, müssen Sie alle Effizienzeinstellungen deaktivieren, da die Verwendung dieser Funktionen keine Einsparungen bietet. Die Kombination von Thick Provisioning mit aktivierter Speicherplatzeffizienz kann zu unerwartetem Verhalten führen, einschließlich Fehlern aufgrund von Speicherplatzout.
- Wenn Daten nicht überschrieben werden, wie etwa bei Backups oder Datenbanktransaktionsprotokollen, können Sie die Effizienz steigern, indem Sie TSSE mit einem niedrigen Kühlzeitraum aktivieren.
- Einige Dateien enthalten möglicherweise eine beträchtliche Menge an nicht komprimierbaren Daten. Ein Beispiel: Wenn die Komprimierung bereits auf Applikationsebene aktiviert ist, werden Dateien verschlüsselt. Wenn eines dieser Szenarien zutrifft, sollten Sie die Komprimierung deaktivieren, um einen effizienteren Betrieb auf anderen Volumes mit komprimierbaren Daten zu ermöglichen.
- Verwenden Sie für Datenbank-Backups nicht sowohl die 32-KB-Komprimierung als auch die Deduplizierung. Siehe Abschnitt Anpassungsfähige Komprimierung Entsprechende Details.

Thin Provisioning

Thin Provisioning für eine Oracle-Datenbank erfordert eine sorgfältige Planung, da im Ergebnis mehr Speicherplatz auf einem Storage-System konfiguriert wird, als unbedingt physisch verfügbar ist. Dieser Aufwand lohnt sich wirklich, denn bei korrekter Umsetzung ergeben sich erhebliche Kosteneinsparungen und Verbesserungen beim Management.

Thin Provisioning ist in vielerlei Form verfügbar und integraler Bestandteil zahlreicher Funktionen von ONTAP für Enterprise-Applikationsumgebungen. Aus diesem Grund steht Thin Provisioning auch eng mit Effizienztechnologien im Zusammenhang: Mithilfe von Effizienzfunktionen können mehr logische Daten gespeichert werden, als dies technisch auf dem Storage-System möglich ist.

Fast jede Verwendung von Snapshots beinhaltet Thin Provisioning. Zum Beispiel umfasst eine typische 10-TB-Datenbank auf NetApp Storage etwa 30 Tage Snapshots. Diese Anordnung führt dazu, dass ca. 10 TB Daten

im aktiven File-System sichtbar sind und 300 TB für Snapshots dediziert. Die insgesamt 310 TB Storage-Kapazität befindet sich in der Regel auf einem Speicherplatz von 12 TB bis 15 TB. Die aktive Datenbank benötigt 10 TB Storage. Die verbleibenden 300 TB an Daten benötigen nur 2 TB bis 5 TB Speicherplatz, da nur die Änderungen an den Originaldaten gespeichert werden.

Das Klonen ist ebenfalls ein Beispiel für Thin Provisioning. Ein großer NetApp Kunde hat 40 Klone einer 80-TB-Datenbank für die Entwicklung erstellt. Wenn alle 40 Entwickler, die diese Klone verwenden, jeden Block in jeder Datendatei übergeschrieben haben, wäre mehr als 3,2 PB Storage erforderlich. In der Praxis sind Umsätze gering und der kollektive Platzbedarf liegt bei näher bei 40 TB, da nur Änderungen auf den Laufwerken gespeichert werden.

Speicherplatzmanagement

Bei Thin Provisioning in einer Applikationsumgebung ist Vorsicht geboten, da sich die Datenänderungsraten unerwartet erhöhen können. Beispielsweise kann der Speicherplatzverbrauch aufgrund von Snapshots schnell ansteigen, wenn Datenbanktabellen neu indiziert werden, oder es werden umfangreiche Patches für VMware Gäste angewendet. Ein falsch platziertes Backup kann in sehr kurzer Zeit große Datenmengen schreiben. Schließlich kann es schwierig sein, einige Anwendungen wiederherzustellen, wenn ein Dateisystem unerwartet über den freien Speicherplatz verfügt.

Glücklicherweise können diese Risiken mit einer sorgfältigen Konfiguration von behoben werden `volume-autogrow` Und `snapshot-autodelete` Richtlinien. Mit diesen Optionen kann ein Benutzer Richtlinien erstellen, die automatisch den von Snapshots belegten Speicherplatz freigeben oder ein Volume erweitern, um zusätzliche Daten aufzunehmen. Es stehen zahlreiche Optionen zur Verfügung, und die Anforderungen variieren je nach Kunde.

Siehe "Dokumentation des Managements von logischem Storage" Für eine vollständige Diskussion dieser Funktionen.

Fraktionale Reservierungen

Die fraktionale Reserve bezieht sich auf das Verhalten einer LUN in einem Volume in Bezug auf die Platzeffizienz. Wenn die Option `fractional-reserve` Ist auf 100 % festgelegt, können alle Daten im Volume mit jedem Datenmuster 100 % Umsatz verzeichnen, ohne Speicherplatz auf dem Volume zu belegen.

Betrachten Sie beispielsweise eine Datenbank auf einer einzigen 250 GB LUN und einem Volume mit 1 TB. Wenn ein Snapshot erstellt wird, würde sofort eine zusätzliche 250GB an Speicherplatz auf dem Volume reserviert werden, um zu garantieren, dass auf dem Volume aus irgendeinem Grund nicht mehr genügend Speicherplatz verfügbar ist. Die Verwendung von fraktionalen Reserven ist im Allgemeinen aufwändig, da es äußerst unwahrscheinlich ist, dass jedes Byte im Datenbank-Volume überschrieben werden müsste. Es gibt keinen Grund, Platz für ein Ereignis zu reservieren, das nie passiert. Wenn ein Kunde jedoch den Speicherplatzverbrauch in einem Storage-System nicht überwachen kann und sicher sein muss, dass der Platz nie knapp wird, wären für die Nutzung von Snapshots 100 % fraktionale Reservierungen erforderlich.

Komprimierung und Deduplizierung

Komprimierung und Deduplizierung sind beide Formen von Thin Provisioning. Beispielsweise kann ein 50 TB Platzbedarf für Daten auf 30 TB komprimiert werden, was zu Einsparungen von 20 TB führt. Um die Komprimierung nutzen zu können, müssen einige dieser 20 TB für andere Daten verwendet werden. Alternativ muss das Storage-System mit weniger als 50 TB erworben werden. Das Ergebnis sind Speicherung von mehr Daten als technisch auf dem Speichersystem verfügbar ist. Aus Sicht der Daten gibt es 50 TB an Daten, obwohl diese auf den Laufwerken nur 30 TB belegen.

Es besteht immer die Möglichkeit, dass sich die Komprimierbarkeit eines Datensatzes ändert. Dies würde zu einem erhöhten Verbrauch an echtem Speicherplatz führen. Dieser Anstieg des Verbrauchs bedeutet, dass die

Komprimierung wie bei anderen Thin Provisioning-Methoden zur Überwachung und Nutzung gemanagt werden muss `volume-autogrow` Und `snapshot-autodelete`.

Die Komprimierung und Deduplizierung werden im Abschnitt Link:efficiency.html ausführlicher behandelt

Komprimierung und fraktionale Reservierungen

Komprimierung ist eine Form von Thin Provisioning. Fraktionale Reservierungen beeinflussen die Komprimierung. Ein wichtiger Hinweis: Vor der Snapshot-Erstellung wird Speicherplatz reserviert. Normalerweise ist eine fraktionale Reserve nur wichtig, wenn ein Snapshot vorhanden ist. Wenn es keinen Snapshot gibt, ist die fraktionale Reserve nicht wichtig. Dies ist bei der Komprimierung nicht der Fall. Wenn eine LUN auf einem Volume mit Komprimierung erstellt wird, behält ONTAP den Speicherplatz bei, um einen Snapshot aufzunehmen. Dieses Verhalten kann während der Konfiguration verwirrend sein, aber es wird erwartet.

Als Beispiel betrachten Sie ein 10GB Volume mit einer 5GB LUN, die bis auf 2,5 GB ohne Snapshots komprimiert wurde. Betrachten wir die beiden folgenden Szenarien:

- Die fraktionale Reserve = 100 ergibt eine Auslastung von 7,5 GB
- Die fraktionale Reserve = 0 ergibt eine Auslastung von 2,5 GB

Das erste Szenario umfasst 2,5 GB Speicherplatzverbrauch für aktuelle Daten und 5 GB Speicherplatz, um 100 % des Umsatzes der Quelle in Erwartung der Snapshot-Nutzung zu berücksichtigen. Das zweite Szenario reserviert keinen zusätzlichen Speicherplatz.

Obwohl diese Situation verwirrend erscheinen mag, ist es unwahrscheinlich, dass sie in der Praxis angetroffen wird. Komprimierung impliziert Thin Provisioning, und Thin Provisioning in einer LUN-Umgebung erfordert nur fraktionale Reservierungen. Es ist immer möglich, dass komprimierte Daten durch eine nicht komprimierbare Funktion überschrieben werden. Aus diesem Grund muss ein Volume für die Komprimierung bereitgestellt werden, um mögliche Einsparungen zu erzielen.

NetApp empfiehlt die folgenden Reservekonfigurationen:

- Einstellen `fractional-reserve` Auf 0, wenn die grundlegende Kapazitätsüberwachung zusammen mit eingerichtet ist `volume-autogrow` Und `snapshot-autodelete`.
- Einstellen `fractional-reserve` Zu 100, wenn es keine Überwachungsfähigkeit gibt oder wenn es unmöglich ist, unter keinen Umständen Raum abzulassen.

Zuweisung von freiem Speicherplatz und LVM-Speicherplatz

Die Effizienz der dynamischen Bereitstellung aktiver LUNs in einer Dateisystemumgebung kann mit der Zeit durch das Löschen von Daten verloren gehen. Sofern die gelöschten Daten nicht entweder mit Nullen überschrieben werden (siehe auch ASMRU) oder der Speicherplatz durch TRIM/UNMAP-Speicherfreigabe freigegeben wird, belegen die "gelöschten" Daten immer mehr nicht zugewiesenen Speicherplatz im Dateisystem. Darüber hinaus ist Thin Provisioning aktiver LUNs in vielen Datenbankumgebungen nur von begrenztem Nutzen, da die Datendateien bei ihrer Erstellung auf ihre volle Größe initialisiert werden.

Eine sorgfältige Planung der LVM-Konfiguration kann die Effizienz steigern und den Bedarf an Storage-Bereitstellung und LUN-Anpassung minimieren. Wenn eine LVM wie Veritas VxVM oder Oracle ASM verwendet wird, werden die zugrunde liegenden LUNs in Extents unterteilt, die nur bei Bedarf verwendet werden. Wenn beispielsweise ein Datensatz bei einer Größe von 2 TB beginnt, jedoch im Laufe der Zeit bis auf 10 TB anwachsen könnte, könnte dieser Datensatz auf 10 TB an LUNs platziert werden, die über Thin Provisioning in einer LVM-Festplattengruppe organisiert sind. Zum Zeitpunkt der Erstellung würden nur 2 TB

Speicherplatz belegt und zusätzlichen Speicherplatz beanspruchen, wenn Extents zugewiesen werden, um dem Datenwachstum gerecht zu werden. Dieser Prozess ist sicher, solange der Speicherplatz überwacht wird.

ONTAP Failover/Switchover

Kenntnisse über Storage-Takeover- und Switchover-Funktionen sind erforderlich, damit Oracle Datenbankvorgänge nicht durch diese Vorgänge unterbrochen werden. Darüber hinaus können die Argumente für Takeover- und Switchover-Vorgänge die Datenintegrität beeinträchtigen, wenn sie falsch verwendet werden.

- Unter normalen Bedingungen werden eingehende Schreibvorgänge an einen bestimmten Controller synchron mit seinem Partner gespiegelt. In einer NetApp MetroCluster-Umgebung werden Schreibvorgänge auch auf einem Remote Controller gespiegelt. Bis ein Schreibvorgang auf nicht-flüchtigen Medien an allen Standorten gespeichert wird, wird er für die Host-Applikation nicht bestätigt.
- Das Medium, auf dem die Schreibdaten gespeichert sind, wird als nichtflüchtiger Speicher oder NVMEM bezeichnet. Gelegentlich wird dieser Speicher auch als NVRAM (Nonvolatile Random Access Memory) bezeichnet. Er kann als Schreib-Cache verwendet werden, obwohl er als Journal fungiert. Im normalen Betrieb werden die Daten von NVMEM nicht gelesen, sondern nur zum Schutz der Daten bei einem Software- oder Hardwareausfall verwendet. Wenn die Daten auf die Laufwerke geschrieben werden, werden die Daten vom RAM im System und nicht von NVMEM übertragen.
- Während eines Übernahmevorgangs übernimmt ein Node in einem Hochverfügbarkeitspaar (HA) den Betrieb seines Partners. Eine Umschaltung ist im Wesentlichen dieselbe, gilt aber für MetroCluster Konfigurationen, bei denen ein Remote Node die Funktionen eines lokalen Node übernimmt.

Bei routinemäßigen Wartungsvorgängen sollte ein Storage-Takeover- oder Switchover-Vorgang transparent sein. Anders als bei Änderungen der Netzwerkpfade besteht hier eine potenzielle kurze Betriebsunterbrechung. Networking kann jedoch kompliziert sein und es sind leicht Fehler zu machen. NetApp empfiehlt daher dringend, Takeover- und Switchover-Vorgänge sorgfältig zu testen, bevor das Storage-System in Betrieb geht. Nur so können Sie sicherstellen, dass alle Netzwerkpfade korrekt konfiguriert sind. Prüfen Sie in einer SAN-Umgebung die Ausgabe des Befehls sorgfältig `sanlun lun show -p` Um sicherzustellen, dass alle erwarteten primären und sekundären Pfade verfügbar sind.

Bei erzwungener Übernahme oder Umschaltung ist Vorsicht zu beachten. Eine Änderung der Storage-Konfiguration mit diesen Optionen erzwingen, bedeutet, dass der Status des Controllers, dem die Laufwerke gehören, nicht berücksichtigt wird und der alternative Node gewaltsam die Kontrolle über die Laufwerke übernimmt. Ein falscher erzwingen eines Takeover kann zu Datenverlust oder Datenkorruption führen. Das liegt daran, dass durch eine erzwungene Übernahme oder Umschaltung die Inhalte von NVMEM verworfen werden. Nach Abschluss der Übernahme oder Umschaltung bedeutet der Verlust dieser Daten, dass die auf den Laufwerken gespeicherten Daten aus Sicht der Datenbank möglicherweise wieder in einen etwas älteren Zustand zurückgesetzt werden.

Eine erzwungene Übernahme mit einem normalen HA-Paar sollte selten erforderlich sein. In fast allen Ausfallszenarien schaltet ein Node ab und informiert den Partner, sodass eine automatische Ausfallsicherung stattfindet. Es gibt einige Edge-Fälle, beispielsweise einen Rolling Failure, bei dem die Verbindung zwischen den Nodes unterbrochen wird und dann ein Controller verloren geht. Dadurch ist eine erzwungene Übernahme erforderlich. In einer solchen Situation geht die Spiegelung zwischen Nodes vor dem Controller-Ausfall verloren. Das bedeutet, dass der verbleibende Controller nicht mehr über eine Kopie der laufenden Schreibvorgänge verfügt. Das Takeover muss anschließend forciert werden, d. h., dass Daten potenziell verloren gehen.

Dieselbe Logik gilt auch für eine MetroCluster-Umschaltung. Unter normalen Bedingungen ist eine Umschaltung nahezu transparent. Bei einem Ausfall kann es jedoch zu einem Verlust der Verbindung zwischen dem noch intakten Standort und dem Notfallstandort kommen. Aus Sicht des verbleibenden

Standorts könnte das Problem lediglich eine Unterbrechung der Verbindung zwischen den Standorten sein, wobei der ursprüngliche Standort möglicherweise noch die Daten verarbeitet. Wenn ein Node den Status des primären Controllers nicht überprüfen kann, ist nur eine erzwungene Umschaltung möglich.

NetApp empfiehlt die folgenden Vorsichtsmaßnahmen zu ergreifen:

- Seien Sie vorsichtig, damit Sie nicht versehentlich eine Übernahme oder Umschaltung erzwingen. Normalerweise sollte das Erzwingen nicht erforderlich sein, und das Erzwingen der Änderung kann zu Datenverlust führen.
- Wenn eine erzwungene Übernahme oder Umschaltung erforderlich ist, stellen Sie sicher, dass die Applikationen heruntergefahren, alle Filesysteme getrennt und die Volume-Gruppen des Logical Volume Manager (LVM) unterschiedlich sind. ASM-Diskgroups müssen abgehängt werden.
- Sollte eine erzwungene MetroCluster-Umschaltung stattfinden, sollte der ausgefallene Node von allen verbleibenden Storage-Ressourcen abgetrennt werden. Weitere Informationen zur entsprechenden Version von ONTAP finden Sie im MetroCluster-Management- und Disaster-Recovery-Leitfaden.

MetroCluster und mehrere Aggregate

MetroCluster ist eine Technologie für die synchrone Replizierung, die bei einer Unterbrechung der Verbindung zum asynchronen Modus wechselt. Dies ist die häufigste Anforderung von Kunden, da durch die garantierte synchrone Replizierung eine Unterbrechung der Standortkonnektivität zu einem vollständigen Stillstand der Datenbank-I/O führt und die Datenbank außer Betrieb genommen wird.

Mit MetroCluster synchronisieren sie Aggregate nach der Wiederherstellung der Konnektivität schnell neu. Im Gegensatz zu anderen Storage-Technologien sollte bei MetroCluster nach einem Standortausfall nie eine vollständige Respiegelung erforderlich sein. Es müssen nur Delta-Änderungen versendet werden.

Bei Datensätzen, die sich über Aggregate verteilen, besteht das geringe Risiko, dass bei einem rollierenden Disaster-Szenario zusätzliche Schritte zur Daten-Recovery erforderlich wären. Insbesondere, wenn (a) die Verbindung zwischen den Standorten unterbrochen wird, (b) die Konnektivität wiederhergestellt wird, (c) die Aggregate einen Zustand erreichen, in dem einige synchronisiert werden und andere nicht, und dann (d) der primäre Standort verloren geht, ist das Ergebnis ein noch existender Standort, an dem die Aggregate nicht miteinander synchronisiert werden. In diesem Fall werden Teile des Datensatzes miteinander synchronisiert. Ohne Recovery können Applikationen, Datenbanken oder Datastores nicht mehr angezeigt werden. Wenn ein Datensatz über mehrere Aggregate verteilt ist, empfiehlt NetApp dringend, Snapshot-basierte Backups mit einem der vielen verfügbaren Tools einzusetzen, um in diesem ungewöhnlichen Szenario eine schnelle Wiederherstellbarkeit zu überprüfen.

ONTAP Konfiguration auf ASA r2-Systemen

RAID

RAID bezeichnet die Verwendung von Paritäts-basierter Redundanz zum Schutz von Daten vor Festplattenausfällen. ASA r2 nutzt die gleichen ONTAP RAID-Technologien wie AFF und FAS -Systeme und gewährleistet so einen robusten Schutz vor dem Ausfall mehrerer Festplatten.

ONTAP führt die RAID-Konfiguration für ASA r2-Systeme automatisch durch. Dies ist ein Kernbestandteil der vereinfachten Speicherverwaltung, die mit der ASA r2-Persönlichkeit eingeführt wurde.

Wichtige Details zur automatischen RAID-Konfiguration auf ASA r2 sind:

- Storage Availability Zones (SAZ): Anstatt herkömmliche Aggregate und RAID-Gruppen manuell zu verwalten, verwendet ASA r2 Storage Availability Zones (SAZs). Hierbei handelt es sich um gemeinsam genutzte, RAID-geschützte Festplattenpools für ein HA-Paar, bei dem beide Knoten vollen Zugriff auf denselben Speicher haben.
- Automatische Platzierung: Wenn eine Speichereinheit (LUN oder NVMe-Namespace) erstellt wird, erstellt ONTAP automatisch ein Volume innerhalb der SAZ und platziert es für eine optimale Leistungs- und Kapazitätsbalance.
- Keine manuelle Aggregatverwaltung: Herkömmliche Aggregat- und RAID-Gruppenverwaltungsbefehle werden auf ASA r2 nicht unterstützt. Dadurch entfällt für Administratoren die Notwendigkeit, RAID-Gruppengrößen, Paritätsplatten oder Knotenzuweisungen manuell zu planen.
- Vereinfachte Bereitstellung: Die Bereitstellung erfolgt über den System Manager oder vereinfachte CLI-Befehle, die sich auf die Speichereinheiten und nicht auf das zugrunde liegende physische RAID-Layout konzentrieren.
- Workload-Neuverteilung: Ab der Version 2025 (ONTAP 9.17.1) gleicht ONTAP die Workloads zwischen den Knoten im HA-Paar automatisch neu aus, um sicherzustellen, dass Leistung und Speicherplatznutzung ohne manuelles Eingreifen im Gleichgewicht bleiben.

ASA r2 verwendet automatisch die Standard-RAID-Technologien von ONTAP: RAID DP für die meisten Konfigurationen und RAID-TEC für sehr große SSD-Pools. Dadurch entfällt die Notwendigkeit einer manuellen RAID-Auswahl. Diese auf Parität basierenden RAID-Level bieten eine bessere Speichereffizienz und Zuverlässigkeit als die Spiegelung, die in älteren Oracle-Best Practices oft empfohlen wird, aber für ASA r2 nicht relevant ist. ONTAP umgeht den bei RAID üblichen Schreibverlust durch die WAFL Integration und gewährleistet so eine optimale Leistung für Oracle-Workloads wie Redo-Logging und zufällige Datendateischreibvorgänge. In Kombination mit automatisiertem RAID-Management und Storage Availability Zones bietet ASA r2 hohe Verfügbarkeit und Schutz auf Enterprise-Niveau für Oracle-Datenbanken.

Kapazitätsmanagement

Für das Management von Datenbanken oder anderen Enterprise-Applikationen mit vorhersehbarem, leicht verwaltbarem, hochperformantem Enterprise-Storage ist auf den Laufwerken freier Speicherplatz für das Daten- und Metadaten-Management erforderlich. Die Menge des freien Speicherplatzes hängt vom Typ des verwendeten Laufwerks und von den Geschäftsprozessen ab.

ASA r2 verwendet Storage Availability Zones (SAZ) anstelle von Aggregaten, aber das Prinzip bleibt dasselbe: Freier Speicherplatz umfasst jegliche physische Kapazität, die nicht von tatsächlichen Daten, Snapshots oder System-Overhead verbraucht wird. Auch Thin Provisioning muss berücksichtigt werden – logische Zuweisungen spiegeln nicht die tatsächliche physische Nutzung wider.

Die NetApp Empfehlungen für ASA r2-Speichersysteme, die für Unternehmensanwendungen verwendet werden, lauten wie folgt:

SSD-Pools in ASA r2-Systemen

* NetApp empfiehlt*, in ASA r2-Umgebungen mindestens 10 % freien physischen Speicherplatz vorzuhalten. Diese Richtlinie gilt für SSD-basierte Speicherpools, die von ASA r2-Systemen verwendet werden, und umfasst den gesamten ungenutzten Speicherplatz innerhalb der SAZ und der Speichereinheiten. Logischer Speicherplatz ist unwichtig; der Fokus liegt auf dem tatsächlich verfügbaren freien physischen Speicherplatz für die Datenspeicherung.

Während ASA r2 eine hohe Auslastung ohne Leistungseinbußen aufrechterhalten kann, erhöht der Betrieb nahe der Vollauslastung das Risiko von Platzmangel und den administrativen Aufwand bei der Speichererweiterung. Eine Auslastung von über 90 % beeinträchtigt möglicherweise nicht die Leistung, kann aber die Verwaltung erschweren und die Bereitstellung zusätzlicher Laufwerke verzögern.

ASA r2-Systeme unterstützen Speichereinheiten bis zu 128 TB und SAZ-Größen bis zu 2 PB pro HA-Paar, wobei ONTAP die Kapazität automatisch über die Knoten verteilt. Die Überwachung der Auslastung auf Cluster-, SAZ- und Speichereinheitsebene ist unerlässlich, um ausreichend freien Speicherplatz für Snapshots, Thin-Provisioning-Workloads und zukünftiges Wachstum zu gewährleisten. Wenn die Kapazität kritische Schwellenwerte erreicht (ca. 90 % Auslastung), sollten zusätzliche SSDs in Gruppen (mindestens sechs Laufwerke) hinzugefügt werden, um Leistung und Ausfallsicherheit zu gewährleisten.

Storage Virtual Machines

Die Oracle-Datenbankspeicherverwaltung auf ASA r2-Systemen ist ebenfalls auf einer Storage Virtual Machine (SVM) zentralisiert, die in der ONTAP CLI als vserver bezeichnet wird.

Eine SVM ist die grundlegende Einheit für Speicherbereitstellung und -sicherheit in ONTAP, vergleichbar mit einer Gast-VM auf einem VMware ESX-Server. Bei der Erstinstallation von ONTAP auf ASA r2 verfügt es über keine Datenbereitstellungsfunktionen, bis eine SVM erstellt wird. Die SVM definiert die Persönlichkeit und die Datendienste für die SAN-Umgebung.

ASA r2-Systeme verwenden eine SAN-only ONTAP Persönlichkeit, die auf die Unterstützung von Blockprotokollen (FC, iSCSI, NVMe/FC, NVMe/TCP) optimiert ist und NAS-bezogene Funktionen entfernt. Dies vereinfacht die Verwaltung und stellt sicher, dass alle SVM-Konfigurationen für SAN-Workloads optimiert sind. Im Gegensatz zu AFF/ FAS -Systemen bietet ASA r2 keine Optionen für NAS-Dienste wie Home-Verzeichnisse oder NFS-Freigaben.

Wenn ein Cluster erstellt wird, stellt ASA r2 automatisch eine Standard-Daten-SVM mit dem Namen svm1 bereit, bei der SAN-Protokolle aktiviert sind. Diese SVM ist für Block-Storage-Operationen bereit, ohne dass eine manuelle Konfiguration der Protokolldienste erforderlich ist. Standardmäßig unterstützen die IP-Daten-LIFs in dieser SVM die Protokolle iSCSI und NVMe/TCP und verwenden die Dienstrichtlinie default-data-blocks, was die Ersteinrichtung für SAN-Workloads vereinfacht. Administratoren können später zusätzliche SVMs erstellen oder LIF-Konfigurationen an die Anforderungen von Leistung, Sicherheit oder Mandantenfähigkeit anpassen.

> Logische Schnittstellen (LIFs) für SAN-Protokolle sollten auf Basis von Leistungs- und Verfügbarkeitsanforderungen entworfen werden. ASA r2 unterstützt iSCSI-, FC- und NVMe-LIFs. Beachten Sie jedoch, dass das automatische iSCSI-LIF-Failover nicht standardmäßig aktiviert ist, da ASA r2 für NVMe- und SCSI-Hosts ein gemeinsames Netzwerk verwendet. Um ein automatisches Failover zu aktivieren, erstellen Sie "iSCSI-only LIFs"Die

SVMs

Wie bei anderen ONTAP Plattformen gibt es keine offizielle Best Practice für die Anzahl der zu erstellenden SVMs; die Entscheidung hängt von den Management- und Sicherheitsanforderungen ab.

Die meisten Kunden betreiben eine einzige primäre SVM für den täglichen Betrieb und erstellen zusätzliche SVMs für spezielle Anforderungen, wie zum Beispiel:

- Eine dedizierte SVM für eine kritische Geschäftsdatenbank, die von einem Spezialistenteam verwaltet wird.
- Eine SVM für eine Entwicklungsgruppe mit delegierter administrativer Kontrolle

- Eine SVM für sensible Daten, die einen eingeschränkten administrativen Zugriff erfordern

In Multi-Tenant-Umgebungen kann jedem Mandanten eine dedizierte SVM zugewiesen werden. Die Begrenzung der Anzahl von SVMs und LIFs pro Cluster, HA-Paar und Knoten hängt vom verwendeten Protokoll, dem Knotenmodell und der Version von ONTAP ab. Konsultieren Sie die "NetApp Hardware Universe" für diese Grenzwerte.

ASA r2 unterstützt ab ONTAP 9.18.1 bis zu 256 SVMs pro Cluster und pro HA-Paar (zuvor 32 in früheren Versionen).

Leistungsmanagement mit ONTAP QoS auf ASA r2-Systemen

Für die sichere und effiziente Verwaltung mehrerer Oracle-Datenbanken auf ASA r2 ist eine effektive QoS-Strategie erforderlich. Dies ist besonders wichtig, da ASA r2-Systeme All-Flash-SAN-Plattformen sind, die für extrem hohe Leistung und Workload-Konsolidierung entwickelt wurden.

Bereits eine relativ geringe Anzahl von SSDs kann selbst die leistungsstärksten Controller auslasten, daher sind QoS-Steuerungen unerlässlich, um eine vorhersehbare Leistung über verschiedene Arbeitslasten hinweg zu gewährleisten. Zum Vergleich: ASA r2-Systeme wie die ASA A1K oder A90 erreichen Hunderttausende bis über eine Million IOPS bei Latenzzeiten im Submillisekundenbereich. Nur sehr wenige einzelne Arbeitslasten würden dieses Leistungsniveau ausschöpfen, daher erfordert die volle Auslastung typischerweise das Hosting mehrerer Datenbanken oder Anwendungen. Um dies sicher zu gewährleisten, sind QoS-Richtlinien erforderlich, um Ressourcenkonflikte zu vermeiden.

ONTAP QoS auf ASA r2 funktioniert genauso wie auf AFF/ FAS -Systemen, mit zwei primären Steuerungsarten: IOPS und Bandbreite. QoS-Steuerungen können auf SVMs und LUNs angewendet werden.

IOPS QoS

IOPS-basierte QoS begrenzt die Gesamt-IOPS für eine gegebene Ressource. In ASA r2 können QoS-Richtlinien auf SVM-Ebene und auf einzelne Speicherobjekte wie LUNs angewendet werden. Wenn eine Arbeitslast ihr IOPS-Limit erreicht, werden zusätzliche E/A-Anforderungen in eine Warteschlange für Tokens gestellt, was zu Latenz führt. Dies ist ein erwartbares Verhalten und verhindert, dass eine einzelne Arbeitslast die Systemressourcen monopolisiert.

Bei der Anwendung von QoS-Kontrollen auf Datenbank-Transaktions-/Redo-Log-Daten ist Vorsicht geboten. Diese Arbeitslasten treten stoßweise auf, und ein QoS-Limit, das für die durchschnittliche Aktivität angemessen erscheint, kann für Spitzenlasten zu niedrig sein und zu erheblichen Leistungsproblemen führen. Im Allgemeinen sollten Redo- und Archivierungsprotokollierung nicht durch QoS eingeschränkt werden.

Bandbreiten QoS

Bandbreitenbasierte QoS begrenzt den Durchsatz in Mbit/s. Dies ist nützlich, wenn Arbeitslasten große Blocklese- oder Schreibvorgänge durchführen, wie z. B. vollständige Tabellenscans oder Sicherungsvorgänge, die eine erhebliche Bandbreite, aber relativ wenige IOPS verbrauchen. Durch die Kombination von IOPS- und Bandbreitenbegrenzungen lässt sich eine feinere Steuerung erreichen.

Minimum/garantierte QoS

Mindest-QoS-Richtlinien reservieren Leistung für kritische Arbeitslasten. In einer gemischten Umgebung mit

Produktions- und Entwicklungsdatenbanken sollte beispielsweise für Entwicklungs-Workloads die maximale QoS und für Produktions-Workloads die minimale QoS angewendet werden, um eine vorhersehbare Leistung zu gewährleisten.

Anpassungsfähige QoS

Adaptive QoS passt die Grenzwerte basierend auf der Größe des Speicherobjekts an. Obwohl es selten für Datenbanken verwendet wird (da die Größe nicht mit den Leistungsanforderungen korreliert), kann es für Virtualisierungs-Workloads nützlich sein, bei denen die Leistungsanforderungen mit der Kapazität skalieren.

Effizienz

Die Funktionen zur Speicherplatzoptimierung von ONTAP werden vollständig unterstützt und für ASA r2-Systeme optimiert. In fast allen Fällen ist es am besten, die Standardeinstellungen beizubehalten und alle Effizienzfunktionen zu aktivieren.

Bei ASA r2-Systemen handelt es sich um All-Flash-SAN-Plattformen, daher sind Effizienztechnologien wie Komprimierung, Verdichtung und Deduplizierung entscheidend für die Maximierung der nutzbaren Kapazität und die Senkung der Kosten.

Komprimierung

Die Komprimierung reduziert den Speicherplatzbedarf durch die Kodierung von Mustern in den Daten. Bei SSD-basierten ASA r2-Systemen führt die Komprimierung zu erheblichen Einsparungen, da Flash-Speicher die Notwendigkeit einer Überdimensionierung für die Leistung überflüssig macht. Die adaptive Komprimierung von ONTAP ist standardmäßig aktiviert und wurde gründlich mit Unternehmens-Workloads, einschließlich Oracle-Datenbanken, getestet, ohne dass messbare Auswirkungen auf die Leistung festgestellt wurden – selbst in Umgebungen, in denen die Latenz im Mikrosekundenbereich gemessen wird. In einigen Fällen verbessert sich die Leistung, weil komprimierte Daten weniger Cache-Speicherplatz belegen.

Bei ASA r2-Systemen wird keine temperaturabhängige Speichereffizienz (TSSE) angewendet. Bei ASA r2-Systemen basiert die Komprimierung nicht auf der Unterscheidung zwischen häufig abgerufenen (Hot Data) und selten abgerufenen (Cold Data). Die Komprimierung beginnt, ohne darauf zu warten, dass die Daten kalt werden.

Anpassungsfähige Komprimierung

Die adaptive Komprimierung verwendet standardmäßig eine Blockgröße von 8 KB, was der Blockgröße entspricht, die üblicherweise von relationalen Datenbanken verwendet wird. Größere Blockgrößen (16 KB oder 32 KB) können die Effizienz für sequentielle Daten wie Transaktionsprotokolle oder Backups verbessern, sollten aber bei aktiven Datenbanken mit Vorsicht eingesetzt werden, um den Overhead beim Überschreiben zu vermeiden.

Die Blockgröße kann für ruhende Dateien wie Protokolle oder Sicherungskopien auf bis zu 32 KB erhöht werden. Bevor Sie die Standardeinstellungen ändern, konsultieren Sie die NetApp Anleitung.

Verwenden Sie für Streaming-Backups keine 32-KB-Komprimierung mit Deduplizierung. Verwenden Sie eine 8-KB-Komprimierung, um die Effizienz der Deduplizierung aufrechtzuerhalten.

Kompressionsausrichtung

Die Ausrichtung der Komprimierung ist bei zufälligen Überschreibungen wichtig. Stellen Sie sicher, dass der LUN-Typ, der Partitionsoffset (Vielfaches von 8 KB) und die Dateisystemblockgröße an die Datenbankblockgröße angepasst sind. Sequenzielle Daten wie Backups oder Protokolle erfordern keine Ausrichtungsüberlegungen.

Data-Compaction

Die Kompaktierung ergänzt die Komprimierung, indem sie es ermöglicht, dass mehrere komprimierte Blöcke denselben physischen Block teilen. Wenn beispielsweise ein 8-KB-Block auf 1 KB komprimiert wird, stellt die Komprimierung sicher, dass der verbleibende Speicherplatz nicht verschwendet wird. Diese Funktion ist integriert und führt nicht zu Leistungseinbußen.

Deduplizierung

Durch die Deduplizierung werden doppelte Datenblöcke aus verschiedenen Datensätzen entfernt. Während Oracle-Datenbanken aufgrund eindeutiger Blockheader und -trailer typischerweise nur minimale Einsparungen bei der Deduplizierung erzielen, kann die ONTAP Deduplizierung dennoch Speicherplatz aus Nullblöcken und sich wiederholenden Mustern zurückgewinnen.

Effizienz und Thin Provisioning

ASA r2-Systeme verwenden standardmäßig Thin Provisioning. Effizienzfunktionen ergänzen Thin Provisioning, um die nutzbare Kapazität zu maximieren.

> Speichereinheiten werden auf ASA r2-Speichersystemen immer dünn bereitgestellt. Thick Provisioning wird nicht unterstützt.

QuickAssist-Technologie (QAT)

Auf NetApp ASA r2-Plattformen sorgt die Intel QuickAssist Technology (QAT) für eine hardwarebeschleunigte Effizienzsteigerung, die sich deutlich von der softwarebasierten Temperature-Sensitive Storage Efficiency (TSSE) ohne QAT unterscheidet.

QAT mit Hardwarebeschleunigung:

- Entlastet die CPU-Kerne von Komprimierungs- und Verschlüsselungsaufgaben.
- Ermöglicht die sofortige, effiziente Verarbeitung sowohl von häufig abgerufenen als auch von selten abgerufenen Daten.
- Reduziert die CPU-Auslastung erheblich.
- Bietet höheren Durchsatz und geringere Latenz.
- Verbessert die Skalierbarkeit für leistungssensible Vorgänge wie TLS- und VPN-Verschlüsselung.

TSSE ohne QAT:

- Setzt für einen effizienten Betrieb auf CPU-gesteuerte Prozesse.
- Die Effizienz wird erst nach einer Verzögerung auf kalte Daten angewendet.
- Verbraucht mehr CPU-Ressourcen.
- Schränkt die Gesamtleistung im Vergleich zu QAT-beschleunigten Systemen ein.

Moderne ASA r2-Systeme bieten daher eine schnellere, hardwarebeschleunigte Effizienz und eine bessere Systemauslastung als ältere TSSE-Plattformen.

Bewährte Verfahren zur Effizienzsteigerung für ASA r2

NetApp empfiehlt Folgendes:

ASA r2 Standardeinstellungen

Auf ONTAP Systemen, die auf ASA r2-Systemen laufen, erstellte Speichereinheiten werden Thin Provisioning mit allen standardmäßig aktivierten Inline-Effizienzfunktionen, einschließlich Komprimierung, Kompaktierung und Deduplizierung, durchgeführt. Obwohl Oracle-Datenbanken im Allgemeinen nicht wesentlich von der Deduplizierung profitieren und nicht komprimierbare Daten enthalten können, sind diese Standardeinstellungen für fast alle Arbeitslasten angemessen. ONTAP ist so konzipiert, dass es alle Arten von Daten und E/A-Mustern effizient verarbeitet, unabhängig davon, ob dadurch Kosteneinsparungen erzielt werden. Standardeinstellungen sollten nur dann geändert werden, wenn die Gründe dafür vollständig verstanden werden und ein klarer Vorteil durch die Abweichung besteht.

Allgemeine Empfehlungen

- Komprimierung für verschlüsselte oder anwendungskomprimierte Daten deaktivieren: Wenn Dateien bereits auf Anwendungsebene komprimiert oder verschlüsselt sind, deaktivieren Sie die Komprimierung, um die Leistung zu optimieren und einen effizienteren Betrieb auf anderen Speichereinheiten zu ermöglichen.
- Vermeiden Sie die Kombination von großen Komprimierungsblöcken mit Deduplizierung: Verwenden Sie für Datenbanksicherungen nicht sowohl 32-KB-Komprimierung als auch Deduplizierung. Für Streaming-Backups sollte eine 8-KB-Komprimierung verwendet werden, um die Effizienz der Deduplizierung aufrechtzuerhalten.
- Effizienzsteigerungen überwachen: Nutzen Sie die ONTAP Tools (System Manager, Active IQ), um die tatsächlichen Platzeinsparungen zu verfolgen und die Richtlinien gegebenenfalls anzupassen.

Thin Provisioning

Thin Provisioning für eine Oracle-Datenbank auf ASA r2 erfordert sorgfältige Planung, da dabei mehr logischer Speicherplatz konfiguriert werden muss, als physisch verfügbar ist. Bei korrekter Implementierung bietet Thin Provisioning erhebliche Kosteneinsparungen und eine verbesserte Verwaltbarkeit.

Thin Provisioning ist integraler Bestandteil von ASA r2 und eng mit ONTAP Effizienztechnologien verwandt, da beide die Speicherung von mehr logischen Daten ermöglichen, als die physische Kapazität des Systems zulässt. ASA r2-Systeme sind reine SAN-Systeme, und Thin Provisioning gilt für Speichereinheiten und LUNs innerhalb von Storage Availability Zones (SAZ).

ASA r2-Speichereinheiten sind standardmäßig Thin Provisioning-fähig.

Nahezu jede Verwendung von Snapshots beinhaltet Thin Provisioning. Eine typische 10 TiB große Datenbank mit 30 Tagen Snapshots erscheint beispielsweise als 310 TiB logische Daten, aber es werden nur 12 bis 15 TiB physischer Speicherplatz belegt, da in den Snapshots nur geänderte Blöcke gespeichert werden.

Ähnlich verhält es sich mit dem Klonen, das eine weitere Form der schlanken Bereitstellung darstellt. Eine Entwicklungsumgebung mit 40 Klonen einer 80 TiB großen Datenbank würde bei vollständiger Speicherung 3,2 PiB Speicherplatz benötigen, verbraucht in der Praxis aber weit weniger, da nur die Änderungen

gespeichert werden.

Speicherplatzmanagement

Bei Thin Provisioning in einer Anwendungsumgebung ist Vorsicht geboten, da die Datenänderungsraten unerwartet ansteigen können. Beispielsweise kann der Speicherplatzverbrauch aufgrund von Snapshots rapide ansteigen, wenn Datenbanktabellen neu indiziert werden oder umfangreiche Patches auf VMware-Gastsysteme angewendet werden. Ein verlegtes Backup kann in kürzester Zeit eine große Datenmenge schreiben. Schließlich kann es schwierig sein, einige Anwendungen wiederherzustellen, wenn auf einer LUN unerwartet der freie Speicherplatz ausgeht.

In ASA r2 werden diese Risiken durch **Thin Provisioning**, **proaktive Überwachung** und **LUN-Größenänderungsrichtlinien** gemindert, anstatt durch ONTAP Funktionen wie Volume-Autogrow oder Snapshot-Autodelete. Administratoren sollten:

- Thin Provisioning auf LUNs aktivieren (`space-reserve disabled`) - Dies ist die Standardeinstellung in ASA r2.
- Überwachen Sie die Kapazität mithilfe von System Manager-Warnungen oder API-basierter Automatisierung.
- Nutzen Sie geplante oder skriptbasierte LUN-Größenanpassungen, um dem Wachstum gerecht zu werden.
- Snapshot-Reservierung und automatische Snapshot-Löschung über den System Manager (GUI) konfigurieren.

> Eine sorgfältige Planung der Speicherplatzschwellenwerte und Automatisierungsskripte ist unerlässlich, da ASA r2 weder automatisches Volume-Wachstum noch CLI-gesteuerte Snapshot-Löschung unterstützt.

ASA r2 verwendet keine Fractional-Reserve-Einstellungen, da es sich um eine reine SAN-Architektur handelt, die WAFL-basierte Volume-Optionen abstrahiert. Stattdessen werden Speicherplatzeffizienz und Überschreibungsschutz auf LUN-Ebene verwaltet. Wenn Sie beispielsweise eine 250 GiB große LUN von einer Speichereinheit bereitgestellt haben, verbrauchen Snapshots Speicherplatz basierend auf den tatsächlichen Blockänderungen, anstatt im Voraus eine gleich große Menge Speicherplatz zu reservieren. Dadurch entfällt die Notwendigkeit großer statischer Reservierungen, die in traditionellen ONTAP Umgebungen mit fraktioneller Reserve üblich waren.

> Wenn ein garantierter Überschreibungsschutz erforderlich ist und eine Überwachung nicht möglich ist, sollten Administratoren ausreichend Speicherkapazität bereitstellen und die Snapshot-Reserve entsprechend einstellen. Aufgrund der Bauweise von ASA r2 ist eine Teilreserve jedoch für die meisten Arbeitslasten nicht erforderlich.

Komprimierung und Deduplizierung

Komprimierung und Deduplizierung in ASA r2 sind Technologien zur Speicherplatzoptimierung und keine traditionellen Thin-Provisioning-Mechanismen. Diese Funktionen reduzieren den physischen Speicherbedarf durch die Eliminierung redundanter Daten und die Komprimierung von Blöcken, wodurch mehr logische Daten gespeichert werden können, als die reine Speicherkapazität sonst zulassen würde.

Ein 50 TiB großer Datensatz könnte beispielsweise auf 30 TiB komprimiert werden, wodurch 20 TiB Speicherplatz eingespart werden. Aus Anwendungssicht sind es immer noch 50 TiB Daten, obwohl sie nur 30 TiB auf der Festplatte belegen.

Die Komprimierbarkeit eines Datensatzes kann sich im Laufe der Zeit ändern, was den physischen Speicherplatzbedarf erhöhen kann. Daher müssen Komprimierung und Deduplizierung proaktiv durch Überwachung und Kapazitätsplanung gesteuert werden.

Zuweisung von freiem Speicherplatz und LVM-Speicherplatz

Thin Provisioning in ASA r2-Umgebungen kann mit der Zeit an Effizienz verlieren, wenn gelöschte Blöcke nicht wiederhergestellt werden. Sofern der Speicherplatz nicht mittels TRIM/UNMAP freigegeben oder mit Nullen überschrieben wird (über ASMRU – Automatic Space Management and Reclamation Utility), belegen gelöschte Daten weiterhin physischen Speicherplatz. In vielen Oracle-Datenbankumgebungen bietet Thin Provisioning nur begrenzten Nutzen, da Datendateien typischerweise bereits bei ihrer Erstellung auf ihre volle Größe vorab zugewiesen werden.

Durch sorgfältige Planung der LVM-Konfiguration lassen sich die Effizienz steigern und der Bedarf an Speicherbereitstellung und LUN-Größenänderung minimieren. Bei Verwendung eines LVM wie Veritas VxVM oder Oracle ASM werden die zugrunde liegenden LUNs in Extents unterteilt, die nur bei Bedarf verwendet werden. Wenn ein Datensatz beispielsweise mit einer Größe von 2 TiB beginnt, aber im Laufe der Zeit auf 10 TiB anwachsen kann, könnte dieser Datensatz auf 10 TiB Thin-Provisioned-LUNs platziert werden, die in einer LVM-Diskgroup organisiert sind. Es würde zum Zeitpunkt seiner Erstellung nur 2 TiB Speicherplatz belegen und erst dann zusätzlichen Speicherplatz benötigen, wenn zur Aufnahme des Datenwachstums Extents zugewiesen werden. Dieser Prozess ist sicher, solange der Raum überwacht wird.

ONTAP Failover

Kenntnisse über Speicherübernahmefunktionen sind erforderlich, um sicherzustellen, dass der Betrieb der Oracle-Datenbank während dieser Vorgänge nicht unterbrochen wird. Darüber hinaus können die bei Übernahmen verwendeten Argumente die Datenintegrität beeinträchtigen, wenn sie falsch eingesetzt werden.

Unter normalen Bedingungen werden eingehende Schreibvorgänge an einen bestimmten Controller synchron an seinen HA-Partner gespiegelt. In einer ASA r2-Umgebung mit SnapMirror Active Sync (SM-as) werden Schreibvorgänge auch auf einen Remote-Controller am sekundären Standort gespiegelt. Solange ein Schreibvorgang nicht an allen Speicherorten in nichtflüchtigen Medien gespeichert ist, wird er der Host-Anwendung nicht bestätigt.

Das Medium, in dem die Schreibdaten gespeichert werden, wird als nichtflüchtiger Speicher (NVMEM) bezeichnet. Manchmal wird es auch als nichtflüchtiger Direktzugriffsspeicher (NVRAM) bezeichnet und kann eher als Schreibjournal denn als Cache betrachtet werden. Im Normalbetrieb werden keine Daten aus dem NVMEM gelesen; es dient lediglich dem Schutz der Daten im Falle eines Software- oder Hardwareausfalls. Beim Schreiben von Daten auf Laufwerke werden die Daten aus dem System-RAM und nicht aus dem NVMEM übertragen.

Bei einer Übernahmeoperation übernimmt ein Knoten in einem HA-Paar die Operationen von seinem Partner. Bei ASA r2 ist ein Switchover nicht möglich, da MetroCluster nicht unterstützt wird; stattdessen bietet SnapMirror Active Sync Redundanz auf Standortebene. Die Übernahme des Speichers während routinemäßiger Wartungsarbeiten sollte transparent erfolgen, abgesehen von einer kurzen Unterbrechung des Betriebs aufgrund von Änderungen der Netzwerkpfade. Netzwerktechnik kann komplex sein, und Fehler passieren leicht. Deshalb empfiehlt NetApp dringend, Übernahmevorgänge gründlich zu testen, bevor ein Speichersystem in Produktion genommen wird. Nur so kann sichergestellt werden, dass alle Netzwerkpfade korrekt konfiguriert sind. In einer SAN-Umgebung überprüfen Sie den Pfadstatus mit dem Befehl `sanlun lun show -p` oder die systemeigenen Multipathing-Tools des Betriebssystems, um sicherzustellen, dass alle erwarteten Pfade verfügbar sind. ASA r2-Systeme bieten alle aktiven optimierten Pfade für LUNs, und Kunden, die NVMe-Namespaces verwenden, sollten auf OS-native Tools zurückgreifen, da NVMe-Pfade nicht von

sanlun abgedeckt werden.

Bei der Durchführung einer Zwangsübernahme ist Vorsicht geboten. Eine erzwungene Änderung der Speicherkonfiguration bedeutet, dass der Zustand des Controllers, dem die Laufwerke gehören, ignoriert wird und der alternative Knoten die Kontrolle über die Laufwerke zwangsweise übernimmt. Eine fehlerhafte erzwungene Übernahme kann zu Datenverlust oder -beschädigung führen, da bei einer erzwungenen Übernahme der Inhalt von NVMEM verworfen werden kann. Nach Abschluss der Übernahme bedeutet der Verlust dieser Daten, dass die auf den Laufwerken gespeicherten Daten aus Sicht der Datenbank möglicherweise in einen etwas älteren Zustand zurückfallen.

Eine erzwungene Übernahme mit einem normalen HA-Paar sollte nur selten erforderlich sein. In nahezu allen Fehlerszenarien schaltet sich ein Knoten ab und informiert den Partner, sodass ein automatisches Failover erfolgt. Es gibt einige Sonderfälle, wie zum Beispiel einen rollierenden Ausfall, bei dem die Verbindung zwischen den Knoten unterbrochen wird und anschließend ein Controller ausfällt, in denen eine erzwungene Übernahme erforderlich ist. In einer solchen Situation geht die Spiegelung zwischen den Knoten vor dem Ausfall des Controllers verloren, was bedeutet, dass der verbleibende Controller keine Kopie der laufenden Schreibvorgänge mehr besitzt. Die Übernahme muss dann erzwungen werden, was bedeutet, dass möglicherweise Daten verloren gehen.

NetApp empfiehlt folgende Vorsichtsmaßnahmen:

- Achten Sie unbedingt darauf, nicht versehentlich eine Übernahme zu erzwingen. Normalerweise sollte ein erzwungener Eingriff nicht erforderlich sein, und ein erzwungener Eingriff kann zu Datenverlust führen.
- Falls eine erzwungene Übernahme erforderlich ist, stellen Sie sicher, dass die Anwendungen heruntergefahren, alle Dateisysteme ausgehängt und die Volume-Gruppen des Logical Volume Manager (LVM) deaktiviert werden. ASM-Diskgruppen müssen ausgehängt werden.
- Im Falle eines Ausfalls auf Standortebene bei Verwendung von SM-as wird auf dem verbleibenden Cluster ein automatisches, ungeplantes Failover mithilfe des ONTAP Mediators eingeleitet, was zu einer kurzen E/A-Pause führt. Anschließend werden die Datenbankübergänge vom verbleibenden Cluster aus fortgesetzt. Weitere Informationen finden Sie unter "SnapMirror Active Sync auf ASA r2-Systemen" für detaillierte Konfigurationsschritte.

Datenbankkonfiguration mit AFF/ FAS -Systemen

Blockgrößen

ONTAP verwendet intern eine variable Blockgröße, d. h. Oracle Datenbanken können mit beliebigen Blockgrößen konfiguriert werden. Allerdings können Blockgrößen des Dateisystems die Performance beeinträchtigen und in einigen Fällen kann eine größere Blockgröße für Wiederherstellungen die Performance verbessern.

Blockgrößen der Datendatei

Einige Betriebssysteme bieten eine Auswahl an Filesystem-Blockgrößen. Bei Filesystemen, die Oracle Datendateien unterstützen, sollte die Blockgröße bei Verwendung der Komprimierung 8 KB betragen. Wenn keine Komprimierung erforderlich ist, kann eine Blockgröße von 8 KB oder 4 KB verwendet werden.

Wenn eine Datendatei auf einem Dateisystem mit einem 512-Byte-Block abgelegt wird, sind falsch

ausgerichtete Dateien möglich. Die LUN und das Filesystem sind möglicherweise basierend auf Empfehlungen von NetApp richtig ausgerichtet, der Datei-I/O wäre jedoch falsch ausgerichtet. Eine solche Fehlausrichtung würde zu schwerwiegenden Leistungsproblemen führen.

Dateisysteme, die Redo-Protokolle unterstützen, müssen eine Blockgröße verwenden, die ein Vielfaches der Redo-Blockgröße ist. Dies erfordert in der Regel, dass sowohl das Redo-Log-Dateisystem als auch das Redo-Protokoll selbst eine Blockgröße von 512 Byte verwenden.

Wiederholen Sie die Blockgrößen

Bei sehr hohen Wiederherstellungsraten ist es möglich, dass 4-KB-Blockgrößen die Performance verbessern, da hohe Wiederherstellungsraten es ermöglichen, I/O in weniger und effizienteren Operationen auszuführen. Wenn Redo-Raten größer als 50 Mbit/s sind, sollten Sie eine 4-KB-Blockgröße testen.

Einige Kundenprobleme wurden mit Datenbanken identifiziert, die Wiederherstellungsprotokolle mit 512-Byte-Blockgröße auf einem Dateisystem mit 4-KB-Blockgröße und vielen sehr kleinen Transaktionen verwenden. Der Mehraufwand, der an der Anwendung mehrerer 512-Byte-Änderungen auf einen einzigen 4-KB-Dateisystemblock beteiligt war, führte zu Performance-Problemen, die behoben wurden, indem das Dateisystem auf eine Blockgröße von 512 Byte geändert wurde.

NetApp empfiehlt, dass Sie die Größe des Redo-Blocks nicht ändern, es sei denn, Sie werden von einem zuständigen Kundensupport oder einer professionellen Serviceorganisation beraten oder die Änderung basiert auf der offiziellen Produktdokumentation.

db_File_Multiblock_read_count

Der `db_file_multiblock_read_count` Der Parameter steuert die maximale Anzahl von Oracle-Datenbankblöcken, die Oracle während sequenzieller I/O-Vorgänge als Einzelvorgang liest

Dieser Parameter wirkt sich jedoch weder auf die Anzahl der Blöcke aus, die Oracle während aller Lesevorgänge liest, noch auf zufälligen I/O. Nur die Blockgröße sequenzieller I/O ist betroffen.

Oracle empfiehlt dem Benutzer, diesen Parameter nicht festzulegen. Dadurch kann die Datenbanksoftware automatisch den optimalen Wert einstellen. Das bedeutet im Allgemeinen, dass dieser Parameter auf einen Wert gesetzt wird, der eine I/O-Größe von 1 MB ergibt. Zum Beispiel würde ein Lesevorgang von 1 MB mit 8-KB-Blöcken 128 Blöcke erfordern, und der Standardwert für diesen Parameter wäre daher 128.

Bei den meisten von NetApp an Kundenstandorten festgestellten Performance-Problemen bei Datenbanken handelt es sich um eine falsche Einstellung für diesen Parameter. Es gab triftige Gründe, diesen Wert mit den Oracle-Versionen 8 und 9 zu ändern. Daher kann der Parameter in vorhanden sein, ohne dass dies bekannt ist `init.ora` Dateien, da die Datenbank auf Oracle 10 und höher aktualisiert wurde. Eine ältere Einstellung von 8 oder 16 beeinträchtigt im Vergleich zu einem Standardwert von 128 erheblich die sequenzielle I/O-Performance.

NetApp empfiehlt die Einstellung `db_file_multiblock_read_count` Der Parameter darf nicht im vorhanden sein `init.ora` Datei: NetApp hat noch nie eine Situation erlebt, in der sich durch die Änderung dieses Parameters die Performance verbesserte. In vielen Fällen wurde jedoch der sequenzielle I/O-Durchsatz deutlich beeinträchtigt.

Filesystemio_options

Der Oracle-Initialisierungsparameter `filesystemio_options` Steuert die Verwendung von asynchronem und direktem I/O.

Entgegen der allgemeinen Auffassung schließen sich asynchroner und direkter I/O nicht gegenseitig aus. NetApp hat festgestellt, dass dieser Parameter in Kundenumgebungen häufig falsch konfiguriert ist und dass diese Fehlkonfiguration direkt für viele Performance-Probleme verantwortlich ist.

Asynchroner I/O bedeutet, dass Oracle-I/O-Vorgänge parallelisiert werden können. Bevor asynchroner I/O auf verschiedenen Betriebssystemen verfügbar war, konfigurierten Anwender zahlreiche dbwriter-Prozesse und änderten die Serverprozesskonfiguration. Bei asynchronem I/O führt das Betriebssystem selbst I/O im Auftrag der Datenbanksoftware hocheffizient und parallel aus. Dieser Prozess gefährdet keine Daten, und kritische Vorgänge wie die Oracle-Wiederherstellungsprotokollierung werden weiterhin synchron ausgeführt.

Direkter I/O umgeht den Puffercache des Betriebssystems. I/O auf einem UNIX-System durchläuft normalerweise den Puffercache des Betriebssystems. Dies ist nützlich für Applikationen, die keinen internen Cache verwalten, aber Oracle hat einen eigenen Puffer-Cache innerhalb des SGA. In fast allen Fällen ist es besser, direkten I/O zu ermöglichen und dem SGA Server-RAM zuzuweisen, anstatt sich auf den Puffercache des Betriebssystems zu verlassen. Oracle SGA nutzt den Speicher effizienter. Wenn I/O den Puffer des Betriebssystems durchläuft, finden weitere Verarbeitungsschritte statt, wodurch die Latenzen erhöht werden. Die erhöhten Latenzen sind besonders bei umfangreichen I/O-Schreibvorgängen spürbar, bei denen eine niedrige Latenz eine wichtige Anforderung ist.

Die Optionen für `filesystemio_options` Sind:

- **Async.** Oracle sendet I/O-Anfragen zur Verarbeitung an das Betriebssystem. Mit diesem Prozess kann Oracle andere Aufgaben ausführen, anstatt auf den I/O-Abschluss zu warten. Dadurch wird die I/O-Parallelisierung erhöht.
- **Directio.** Oracle führt I/O direkt auf physische Dateien aus, anstatt I/O über den Host-BS-Cache zu leiten.
- **None.** Oracle verwendet synchrone und gepufferte I/O. In dieser Konfiguration ist die Wahl zwischen Shared-Server- und dedizierten Server-Prozessen und der Anzahl der dbwriters wichtiger.
- **setall.** Oracle verwendet sowohl asynchrone als auch direkte I/O. In fast allen Fällen, die Verwendung von `setall` Ist optimal.

Der `filesystemio_options` Parameter hat keine Auswirkungen in DNFS- und ASM-Umgebungen. Die Verwendung von DNFS oder ASM führt automatisch zur Verwendung von asynchronem und direktem I/O.

Einige Kunden sind in der Vergangenheit auf Probleme mit asynchronem I/O gestoßen, insbesondere mit früheren Versionen von Red hat Enterprise Linux 4 (RHEL4). Einige veraltete Ratschläge im Internet deuten immer noch darauf hin, dass asynchrone IO aufgrund veraktueller Informationen vermieden wird. Asynchrone I/O-Vorgänge sind auf allen aktuellen Betriebssystemen stabil. Es gibt keinen Grund, es zu deaktivieren, ohne einen bekannten Fehler mit dem Betriebssystem.

Wenn in einer Datenbank gepufferte I/O verwendet wurden, könnte ein Wechsel zu direkten I/O auch eine Änderung der SGA-Größe rechtfertigen. Durch die Deaktivierung gepufferter I/O-Vorgänge werden die Performance-Vorteile eliminiert, die der Host-BS-Cache für die Datenbank bietet. Durch das Hinzufügen von RAM zum SGA wird dieses Problem behoben. Das Nettoergebnis sollte eine Verbesserung der I/O-Performance sein.

Obwohl es fast immer besser ist, RAM für Oracle SGA zu verwenden statt für das Zwischenspeichern von BS-

Puffern, ist es unter Umständen nicht möglich, den besten Wert zu ermitteln. Es könnte beispielsweise besser sein, gepufferten I/O mit sehr kleinen SGA-Größen auf einem Datenbankserver mit vielen intermittierend aktiven Oracle-Instanzen zu verwenden. Diese Anordnung ermöglicht die flexible Nutzung des verbleibenden freien RAM auf dem Betriebssystem durch alle ausgeführten Datenbankinstanzen. Dies ist eine äußerst ungewöhnliche Situation, die jedoch an einigen Kundenstandorten beobachtet wurde.

NetApp empfiehlt Einstellung `filesystemio_options` Bis `setall`, Aber beachten Sie, dass unter bestimmten Umständen der Verlust des Host-Puffer-Caches eine Erhöhung des Oracle SGA erfordern kann.

RAC-Timeouts

Oracle RAC ist ein Clusterware-Produkt mit verschiedenen Arten von internen Heartbeat-Prozessen, die den Zustand des Clusters überwachen.

Die Informationen im "Fehlzählung" Der Abschnitt enthält wichtige Informationen für Oracle RAC-Umgebungen, die Netzwerkspeicher verwenden. In vielen Fällen müssen die standardmäßigen Oracle RAC-Einstellungen geändert werden, um sicherzustellen, dass der RAC-Cluster Netzwerkpfadänderungen und Speicher-Failover-/Switchover-Vorgänge überlebt.

Festplatten-Timeout

Der primäre speicherbezogene RAC-Parameter lautet `disktimeout`. Dieser Parameter steuert den Schwellenwert, innerhalb dessen die Abstimmungsdatei E/A abgeschlossen werden muss. Wenn der `disktimeout` Parameter überschritten wird, dann wird der RAC-Knoten aus dem Cluster entfernt. Der Standardwert für diesen Parameter ist 200. Dieser Wert sollte für standardmäßige Storage-Takeover- und Giveback-Verfahren ausreichen.

NetApp empfiehlt, die RAC-Konfigurationen vor ihrer Inbetriebnahme sorgfältig zu testen, da sich viele Faktoren auf einen Takeover oder Giveback auswirken. Neben der Zeit, die für den Abschluss des Storage-Failovers benötigt wird, ist auch für die Verbreitung der Änderungen des Link Aggregation Control Protocol (LACP) zusätzliche Zeit erforderlich. Darüber hinaus muss die SAN-Multipathing-Software eine I/O-Zeitüberschreitung erkennen und einen alternativen Pfad erneut versuchen. Wenn eine Datenbank extrem aktiv ist, muss eine große Menge an I/O-Vorgängen in die Warteschlange gestellt und erneut versucht werden, bevor die Abstimmungs-E/A-Vorgänge verarbeitet werden.

Wenn ein tatsächlicher Storage Takeover oder Giveback nicht möglich ist, kann der Effekt durch Cable Pull-Tests auf dem Datenbankserver simuliert werden.

NetApp empfiehlt Folgendes:

- Verlassen des `disktimeout` Parameter mit dem Standardwert 200.
- Testen Sie eine RAC-Konfiguration immer gründlich.

Fehlzählung

Der `misscount` Der Parameter wirkt sich normalerweise nur auf den Netzwerk-Heartbeat zwischen RAC-Knoten aus. Die Standardeinstellung ist 30 Sekunden. Wenn sich die Grid-Binärdateien auf einem Storage Array befinden oder das Boot-Laufwerk des Betriebssystems nicht lokal ist, kann dieser Parameter wichtig werden. Dazu gehören Hosts mit Boot-Laufwerken in einem FC-SAN, über NFS gestartete Betriebssysteme und Boot-Laufwerke in Virtualisierungs-Datastores, beispielsweise eine VMDK-Datei.

Wird der Zugriff auf ein Boot-Laufwerk durch eine Storage-Übernahme oder -Rückgabe unterbrochen, kann es sein, dass der Binärstandort des Grid oder das gesamte Betriebssystem vorübergehend nicht verfügbar ist. Die Zeit, die ONTAP bis zum Abschluss des Storage-Vorgangs und zum Ändern von Pfaden und zum Fortsetzen der I/O benötigt, kann größer sein als die `misscount` Schwellenwert. Infolgedessen wird ein Node sofort entfernt, nachdem die Verbindung zur Boot-LUN oder zu den Grid-Binärdateien wiederhergestellt wurde. In den meisten Fällen werden Entfernung und anschließende Neustarts ohne Protokollmeldungen durchgeführt, um den Grund für das Neubooten zu geben. Da nicht alle Konfigurationen betroffen sind, sollten Sie jeden SAN-Booting, NFS-Booting oder Datastore-basierten Host in einer RAC-Umgebung testen, damit RAC stabil bleibt, wenn die Kommunikation zum Startlaufwerk unterbrochen wird.

Bei nicht-lokalen Startlaufwerken oder einem nicht lokalen Dateisystem, das hostet `grid` Binärdateien, die `misscount` Muss entsprechend geändert werden `disktimeout`. Wenn dieser Parameter geändert wird, führen Sie weitere Tests durch, um auch alle Auswirkungen auf das RAC-Verhalten zu identifizieren, z. B. die Node-Failover-Zeit.

NetApp empfiehlt Folgendes:

- Verlassen Sie den `misscount` Parameter mit dem Standardwert 30, sofern keine der folgenden Bedingungen zutrifft:
  - `grid` Binärdateien befinden sich auf einem Network-Attached-Laufwerk, einschließlich NFS-, iSCSI-, FC- und Datastore-basierten Laufwerken.
  - Das Betriebssystem wird über SAN gebootet.
- Prüfen Sie in solchen Fällen die Auswirkungen von Netzwerkunterbrechungen, die den Zugriff auf das Betriebssystem oder beeinträchtigen `GRID_HOME` File-Systeme. In einigen Fällen führen solche Unterbrechungen dazu, dass die Oracle RAC-Daemons abgewürgt werden, was zu einem führen kann `misscount`-Based Timeout und Entfernung. Die Zeitüberschreitung beträgt standardmäßig 27 Sekunden. Dies ist der Wert von `misscount` Minus `reboottime`. In solchen Fällen erhöhen `misscount` Bis 200, um zu entsprechen `disktimeout`.

Datenbankkonfiguration mit ASA r2-Systemen

Blockgrößen

ONTAP verwendet intern eine variable Blockgröße, was bedeutet, dass Oracle-Datenbanken mit jeder gewünschten Blockgröße konfiguriert werden können. Allerdings können sich die Blockgrößen des Dateisystems auf die Leistung auswirken, und in einigen Fällen kann eine größere Redo-Blockgröße die Leistung verbessern.

ASA r2 führt im Vergleich zu AFF/ FAS -Systemen keine Änderungen an den Oracle-Blockgrößenempfehlungen ein. Das Verhalten von ONTAP bleibt auf allen Plattformen einheitlich.

Blockgrößen der Datendatei

Einige Betriebssysteme bieten eine Auswahl an Filesystem-Blockgrößen. Bei Filesystemen, die Oracle Datendateien unterstützen, sollte die Blockgröße bei Verwendung der Komprimierung 8 KB betragen. Wenn keine Komprimierung erforderlich ist, kann eine Blockgröße von 8 KB oder 4 KB verwendet werden.

Wenn eine Datendatei auf einem Dateisystem mit einem 512-Byte-Block abgelegt wird, sind falsch ausgerichtete Dateien möglich. Die LUN und das Filesystem sind möglicherweise basierend auf Empfehlungen

von NetApp richtig ausgerichtet, der Datei-I/O wäre jedoch falsch ausgerichtet. Eine solche Fehlausrichtung würde zu schwerwiegenden Leistungsproblemen führen.

Wiederholen Sie die Blockgrößen

Dateisysteme, die Redo-Protokolle unterstützen, müssen eine Blockgröße verwenden, die ein Vielfaches der Redo-Blockgröße ist. Dies erfordert in der Regel, dass sowohl das Redo-Log-Dateisystem als auch das Redo-Protokoll selbst eine Blockgröße von 512 Byte verwenden.

Bei sehr hohen Wiederherstellungsraten ist es möglich, dass 4-KB-Blockgrößen die Performance verbessern, da hohe Wiederherstellungsraten es ermöglichen, I/O in weniger und effizienteren Operationen auszuführen. Wenn Redo-Raten größer als 50 Mbit/s sind, sollten Sie eine 4-KB-Blockgröße testen.

Einige Kundenprobleme wurden mit Datenbanken identifiziert, die Wiederherstellungsprotokolle mit 512-Byte-Blockgröße auf einem Dateisystem mit 4-KB-Blockgröße und vielen sehr kleinen Transaktionen verwenden. Der Mehraufwand, der an der Anwendung mehrerer 512-Byte-Änderungen auf einen einzigen 4-KB-Dateisystemblock beteiligt war, führte zu Performance-Problemen, die behoben wurden, indem das Dateisystem auf eine Blockgröße von 512 Byte geändert wurde.

NetApp empfiehlt, dass Sie die Größe des Redo-Blocks nicht ändern, es sei denn, Sie werden von einem zuständigen Kundensupport oder einer professionellen Serviceorganisation beraten oder die Änderung basiert auf der offiziellen Produktdokumentation.

db_File_Multiblock_read_count

Der `db_file_multiblock_read_count` Der Parameter steuert die maximale Anzahl von Oracle-Datenbankblöcken, die Oracle während sequenzieller I/O-Vorgänge als Einzelvorgang liest

Im Vergleich zu den AFF/ FAS -Systemen gibt es keine Änderungen bei den Empfehlungen. Das ONTAP Verhalten und die Best Practices von Oracle bleiben auf den Plattformen ASA r2, AFF und FAS identisch.

Dieser Parameter wirkt sich jedoch weder auf die Anzahl der Blöcke aus, die Oracle während aller Lesevorgänge liest, noch auf zufälligen I/O. Nur die Blockgröße sequenzieller I/O ist betroffen.

Oracle empfiehlt dem Benutzer, diesen Parameter nicht festzulegen. Dadurch kann die Datenbanksoftware automatisch den optimalen Wert einstellen. Das bedeutet im Allgemeinen, dass dieser Parameter auf einen Wert gesetzt wird, der eine I/O-Größe von 1 MB ergibt. Zum Beispiel würde ein Lesevorgang von 1 MB mit 8-KB-Blöcken 128 Blöcke erfordern, und der Standardwert für diesen Parameter wäre daher 128.

Bei den meisten von NetApp an Kundenstandorten festgestellten Performance-Problemen bei Datenbanken handelt es sich um eine falsche Einstellung für diesen Parameter. Es gab triftige Gründe, diesen Wert mit den Oracle-Versionen 8 und 9 zu ändern. Daher kann der Parameter in vorhanden sein, ohne dass dies bekannt ist `init.ora` Dateien, da die Datenbank auf Oracle 10 und höher aktualisiert wurde. Eine ältere Einstellung von 8 oder 16 beeinträchtigt im Vergleich zu einem Standardwert von 128 erheblich die sequenzielle I/O-Performance.

NetApp empfiehlt die Einstellung `db_file_multiblock_read_count` Der Parameter darf nicht im vorhanden sein `init.ora` Datei: NetApp hat noch nie eine Situation erlebt, in der sich durch die Änderung dieses Parameters die Performance verbesserte. In vielen Fällen wurde jedoch der sequenzielle I/O-Durchsatz deutlich beeinträchtigt.

Filesystemio_options

Der Oracle-Initialisierungsparameter `filesystemio_options` Steuert die Verwendung von asynchronem und direktem I/O.

Das Verhalten und die Empfehlungen für filesystemio_options auf ASA r2 sind identisch mit denen auf AFF/ FAS -Systemen, da der Parameter Oracle-spezifisch ist und nicht von der Speicherplattform abhängt. ASA r2 verwendet ONTAP wie AFF/ FAS, daher gelten die gleichen Best Practices.

Entgegen der allgemeinen Auffassung schließen sich asynchroner und direkter I/O nicht gegenseitig aus. NetApp hat festgestellt, dass dieser Parameter in Kundenumgebungen häufig falsch konfiguriert ist und dass diese Fehlkonfiguration direkt für viele Performance-Probleme verantwortlich ist.

Asynchroner I/O bedeutet, dass Oracle-I/O-Vorgänge parallelisiert werden können. Bevor asynchroner I/O auf verschiedenen Betriebssystemen verfügbar war, konfigurierten Anwender zahlreiche dbwriter-Prozesse und änderten die Serverprozesskonfiguration. Bei asynchronem I/O führt das Betriebssystem selbst I/O im Auftrag der Datenbanksoftware hocheffizient und parallel aus. Dieser Prozess gefährdet keine Daten, und kritische Vorgänge wie die Oracle-Wiederherstellungsprotokollierung werden weiterhin synchron ausgeführt.

Direkter I/O umgeht den Puffercache des Betriebssystems. I/O auf einem UNIX-System durchläuft normalerweise den Puffercache des Betriebssystems. Dies ist nützlich für Applikationen, die keinen internen Cache verwalten, aber Oracle hat einen eigenen Puffer-Cache innerhalb des SGA. In fast allen Fällen ist es besser, direkten I/O zu ermöglichen und dem SGA Server-RAM zuzuweisen, anstatt sich auf den Puffercache des Betriebssystems zu verlassen. Oracle SGA nutzt den Speicher effizienter. Wenn I/O den Puffer des Betriebssystems durchläuft, finden weitere Verarbeitungsschritte statt, wodurch die Latenzen erhöht werden. Die erhöhten Latenzen sind besonders bei umfangreichen I/O-Schreibvorgängen spürbar, bei denen eine niedrige Latenz eine wichtige Anforderung ist.

Die Optionen für `filesystemio_options` Sind:

- **Async.** Oracle sendet I/O-Anfragen zur Verarbeitung an das Betriebssystem. Mit diesem Prozess kann Oracle andere Aufgaben ausführen, anstatt auf den I/O-Abschluss zu warten. Dadurch wird die I/O-Parallelisierung erhöht.
- **Directio.** Oracle führt I/O direkt auf physische Dateien aus, anstatt I/O über den Host-BS-Cache zu leiten.
- **None.** Oracle verwendet synchrone und gepufferte I/O. In dieser Konfiguration ist die Wahl zwischen Shared-Server- und dedizierten Server-Prozessen und der Anzahl der dbwriters wichtiger.
- **setall.** Oracle verwendet sowohl asynchrone als auch direkte I/O. In fast allen Fällen, die Verwendung von `setall` Ist optimal.

> In ASM-Umgebungen verwendet Oracle automatisch Direct I/O und Asynchronous I/O für ASM-verwaltete Datenträger, so `filesystemio_options` hat keine Auswirkung auf ASM-Disk-Gruppen. Für Nicht-ASM-Bereitstellungen (z. B. Dateisysteme auf SAN-LUNs) gilt Folgendes: `filesystemio_options = setall`Die Dies ermöglicht sowohl asynchrone als auch direkte Ein-/Ausgabe für optimale Leistung.

Bei einigen älteren Betriebssystemen gab es Probleme mit asynchroner Ein-/Ausgabe, was zu veralteten Empfehlungen führte, diese zu vermeiden. Die asynchrone Ein-/Ausgabe ist jedoch stabil und wird von allen gängigen Betriebssystemen vollständig unterstützt. Es gibt keinen Grund, es zu deaktivieren, es sei denn, es wird ein spezifischer Fehler im Betriebssystem identifiziert.

Wenn in einer Datenbank gepufferte I/O verwendet wurden, könnte ein Wechsel zu direkten I/O auch eine Änderung der SGA-Größe rechtfertigen. Durch die Deaktivierung gepufferter I/O-Vorgänge werden die

Performance-Vorteile eliminiert, die der Host-BS-Cache für die Datenbank bietet. Durch das Hinzufügen von RAM zum SGA wird dieses Problem behoben. Das Nettoergebnis sollte eine Verbesserung der I/O-Performance sein.

Obwohl es fast immer besser ist, RAM für Oracle SGA zu verwenden statt für das Zwischenspeichern von BS-Puffern, ist es unter Umständen nicht möglich, den besten Wert zu ermitteln. Es könnte beispielsweise besser sein, gepufferten I/O mit sehr kleinen SGA-Größen auf einem Datenbankserver mit vielen intermittierend aktiven Oracle-Instanzen zu verwenden. Diese Anordnung ermöglicht die flexible Nutzung des verbleibenden freien RAM auf dem Betriebssystem durch alle ausgeführten Datenbankinstanzen. Dies ist eine äußerst ungewöhnliche Situation, die jedoch an einigen Kundenstandorten beobachtet wurde.

* NetApp empfiehlt* Einstellung `filesystemio_options` Zu `setall`Beachten Sie jedoch, dass der Verlust des Host-Puffer-Caches unter bestimmten Umständen eine Vergrößerung der Oracle SGA erforderlich machen kann. ASA r2-Systeme sind für SAN-Workloads mit niedriger Latenz optimiert, daher passt die Verwendung von setall perfekt zum Design von ASA für leistungsstarke Oracle-Bereitstellungen.

RAC-Timeouts

Oracle RAC ist ein Clusterware-Produkt mit verschiedenen Arten von internen Heartbeat-Prozessen, die den Zustand des Clusters überwachen.

ASA r2-Systeme verwenden ONTAP genau wie AFF/ FAS, daher gelten für die Timeout-Parameter von Oracle RAC die gleichen Prinzipien. Es gibt keine ASA-spezifischen Änderungen an den Empfehlungen zu Disk-Timeout oder Fehlzählungen. Allerdings ist ASA r2 für SAN-Workloads und Failover mit geringer Latenz optimiert, weshalb diese Best Practices umso wichtiger sind.

Die Informationen in der "Fehlzählung" Dieser Abschnitt enthält wichtige Informationen für Oracle RAC-Umgebungen mit Netzwerkspeicher, und in vielen Fällen müssen die Standardeinstellungen von Oracle RAC geändert werden, um sicherzustellen, dass der RAC-Cluster Netzwerkpfadänderungen und Speicherausfallvorgänge übersteht.

Festplatten-Timeout

Der primäre speicherbezogene RAC-Parameter lautet `disktimeout`. Dieser Parameter steuert den Schwellenwert, innerhalb dessen die Abstimmungsdatei E/A abgeschlossen werden muss. Wenn der `disktimeout` Parameter überschritten wird, dann wird der RAC-Knoten aus dem Cluster entfernt. Der Standardwert für diesen Parameter ist 200. Dieser Wert sollte für standardmäßige Storage-Takeover- und Giveback-Verfahren ausreichen.

NetApp empfiehlt, die RAC-Konfigurationen vor ihrer Inbetriebnahme sorgfältig zu testen, da sich viele Faktoren auf einen Takeover oder Giveback auswirken. Neben der Zeit, die für den Abschluss des Storage-Failovers benötigt wird, ist auch für die Verbreitung der Änderungen des Link Aggregation Control Protocol (LACP) zusätzliche Zeit erforderlich. Darüber hinaus muss die SAN-Multipathing-Software eine I/O-Zeitüberschreitung erkennen und einen alternativen Pfad erneut versuchen. Wenn eine Datenbank extrem aktiv ist, muss eine große Menge an I/O-Vorgängen in die Warteschlange gestellt und erneut versucht werden, bevor die Abstimmungs-E/A-Vorgänge verarbeitet werden.

Wenn ein tatsächlicher Storage Takeover oder Giveback nicht möglich ist, kann der Effekt durch Cable Pull-Tests auf dem Datenbankserver simuliert werden.

NetApp empfiehlt Folgendes:

- Verlassen des `disktimeout` Parameter mit dem Standardwert 200.
- Testen Sie eine RAC-Konfiguration immer gründlich.

Fehlzählung

Der `misscount` Der Parameter wirkt sich normalerweise nur auf den Netzwerk-Heartbeat zwischen RAC-Knoten aus. Die Standardeinstellung ist 30 Sekunden. Wenn sich die Grid-Binärdateien auf einem Storage Array befinden oder das Boot-Laufwerk des Betriebssystems nicht lokal ist, kann dieser Parameter wichtig werden. Dazu gehören Hosts mit Boot-Laufwerken in einem FC-SAN, über NFS gestartete Betriebssysteme und Boot-Laufwerke in Virtualisierungs-Datastores, beispielsweise eine VMDK-Datei.

Wird der Zugriff auf ein Boot-Laufwerk durch eine Storage-Übernahme oder -Rückgabe unterbrochen, kann es sein, dass der Binärstandort des Grid oder das gesamte Betriebssystem vorübergehend nicht verfügbar ist. Die Zeit, die ONTAP bis zum Abschluss des Storage-Vorgangs und zum Ändern von Pfaden und zum Fortsetzen der I/O benötigt, kann größer sein als die `misscount` Schwellenwert. Infolgedessen wird ein Node sofort entfernt, nachdem die Verbindung zur Boot-LUN oder zu den Grid-Binärdateien wiederhergestellt wurde. In den meisten Fällen werden Entfernung und anschließende Neustarts ohne Protokollmeldungen durchgeführt, um den Grund für das Neubooten zu geben. Da nicht alle Konfigurationen betroffen sind, sollten Sie jeden SAN-Booting, NFS-Booting oder Datastore-basierten Host in einer RAC-Umgebung testen, damit RAC stabil bleibt, wenn die Kommunikation zum Startlaufwerk unterbrochen wird.

Bei nicht-lokalen Startlaufwerken oder einem nicht lokalen Dateisystem, das hostet `grid` Binärdateien, die `misscount` Muss entsprechend geändert werden `disktimeout`. Wenn dieser Parameter geändert wird, führen Sie weitere Tests durch, um auch alle Auswirkungen auf das RAC-Verhalten zu identifizieren, z. B. die Node-Failover-Zeit.

NetApp empfiehlt Folgendes:

- Verlassen Sie den `misscount` Parameter mit dem Standardwert 30, sofern keine der folgenden Bedingungen zutrifft:
  - `grid` Die Binärdateien befinden sich auf einem netzwerkgebundenen Laufwerk, einschließlich iSCSI-, FC- und datenspeicherbasierten Laufwerken.
  - Das Betriebssystem wird über SAN gebootet.
- Prüfen Sie in solchen Fällen die Auswirkungen von Netzwerkunterbrechungen, die den Zugriff auf das Betriebssystem oder beeinträchtigen `GRID_HOME` File-Systeme. In einigen Fällen führen solche Unterbrechungen dazu, dass die Oracle RAC-Daemons abgewürgt werden, was zu einem führen kann `misscount`-Based Timeout und Entfernung. Die Zeitüberschreitung beträgt standardmäßig 27 Sekunden. Dies ist der Wert von `misscount` Minus `reboottime`. In solchen Fällen erhöhen `misscount` Bis 200, um zu entsprechen `disktimeout`.

- Das SAN-optimierte Design der ASA r2 reduziert die Failover-Latenz, aber die Timeouts müssen weiterhin für Netzwerk-Boot oder Grid-Binärdateien angepasst werden.
- Bei erweiterten RAC- oder Active-Active-Setups (z. B. SnapMirror Active Sync) ist die Timeout-Optimierung auch für Zero-RPO-Architekturen unerlässlich.

Hostkonfiguration mit AFF/ FAS -Systemen

AIX

Konfigurationsthemen für Oracle Database auf IBM AIX mit ONTAP.

Gleichzeitige I/O-Vorgänge

Um eine optimale Leistung auf IBM AIX zu erzielen, muss gleichzeitig I/O verwendet werden Ohne gleichzeitige I/O-Vorgänge sind Performance-Einschränkungen wahrscheinlich, weil AIX serialisierte, atomare I/O-Vorgänge durchführt, was einen beträchtlichen Overhead nach sich zieht.

Ursprünglich hat NetApp die Verwendung von empfohlen `cio` Mount-Option, um die Verwendung von gleichzeitigen I/O-Vorgängen auf dem Dateisystem zu erzwingen. Dieser Prozess hatte jedoch Nachteile und ist nicht mehr erforderlich. Seit der Einführung von AIX 5.2 und Oracle 10gR1 kann Oracle auf AIX einzelne Dateien für gleichzeitige I/O-Vorgänge öffnen, anstatt gleichzeitige I/O-Vorgänge auf dem gesamten Dateisystem zu erzwingen.

Die beste Methode für die Aktivierung gleichzeitiger I/O ist, die festzulegen `init.ora` Parameter `filesystemio_options` Bis `setall`. Auf diese Weise kann Oracle spezifische Dateien zur Verwendung mit gleichzeitigen I/O-Vorgängen öffnen

Wird Verwendet `cio` Als Mount-Option erzwingt die Verwendung von gleichzeitigen I/O-Vorgängen, was negative Auswirkungen haben kann. Das Erzwingen von gleichzeitigen I/O-Vorgängen deaktiviert beispielsweise das Vorauslesen auf Dateisystemen, was die Performance für I/O-Vorgänge beeinträchtigen kann, die außerhalb der Oracle-Datenbanksoftware auftreten, z. B. das Kopieren von Dateien und das Durchführen von Bandsicherungen. Darüber hinaus sind Produkte wie Oracle GoldenGate und SAP BR*Tools nicht mit der Verwendung des kompatibel `cio` Mount-Option mit bestimmten Versionen von Oracle.

NetApp empfiehlt Folgendes:

- Verwenden Sie das nicht `cio` Mount-Option auf Filesystem-Ebene. Aktivieren Sie stattdessen Concurrent I/O über `filesystemio_options=setall`.
- Verwenden Sie nur das `cio` Die Mount-Option sollte aktiviert werden, wenn keine Einstellung möglich ist `filesystemio_options=setall`.

Mount-Optionen für AIX NFS

In der folgenden Tabelle sind die AIX NFS-Mount-Optionen für Oracle Single-Instance-Datenbanken aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Steuerdateien Datendateien Wiederherstellungsprotokolle	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`

Dateityp	Mount-Optionen
ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,intr

In der folgenden Tabelle sind die AIX-NFS-Mount-Optionen für RAC aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144
Steuerdateien Datendateien Wiederherstellungsprotokolle	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac
CRS/Voting	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac
Dediziert ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144
Freigegeben ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr

Der Hauptunterschied zwischen Single-Instance- und RAC-Mount-Optionen ist das Hinzufügen von noac Zu den Mount-Optionen. Durch diese Ergänzung wird das Caching des Host-Betriebssystems deaktiviert, wodurch alle Instanzen im RAC Cluster eine konsistente Ansicht des Status der Daten haben.

Obwohl Sie den verwenden cio Mount-Option und der init.ora Parameter filesystemio_options=setall Hat die gleiche Wirkung wie die Deaktivierung des Host-Caching, ist es weiterhin erforderlich, zu verwenden noac. noac Ist für die gemeinsame Nutzung erforderlich ORACLE_HOME Bereitstellung, um die Konsistenz von Dateien wie Oracle-Passwortdateien und zu erleichtern spfile Parameterdateien. Wenn jede Instanz in einem RAC-Cluster über einen dedizierten verfügt ORACLE_HOME, Dann ist dieser Parameter nicht erforderlich.

Mount-Optionen für AIX jfs/jfs2

In der folgenden Tabelle sind die AIX jfs/jfs2-Mount-Optionen aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	Standardwerte
Steuerdateien Datendateien Wiederherstellungsprotokolle	Standardwerte
ORACLE_HOME	Standardwerte

Vor der Verwendung von AIX hdisk Geräte in jeder Umgebung, einschließlich Datenbanken, überprüfen Sie den Parameter queue_depth. Dieser Parameter entspricht nicht der HBA-Warteschlangentiefe, sondern

bezieht sich auf die SCSI-Warteschlangentiefe der einzelnen `hdisk device. Depending on how the LUNs are configured, the value for `queue_depth` Ist möglicherweise zu niedrig für eine gute Leistung. Die Prüfung hat ergeben, dass der optimale Wert 64 ist.

HP-UX ERHÄLTLICH

Konfigurationsthemen für Oracle Database on HP-UX with ONTAP.

HP-UX NFS Mount-Optionen

In der folgenden Tabelle sind die HP-UX-NFS-Mount-Optionen für eine einzelne Instanz aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Kontrolldateien Datendateien Wiederherstellungsprotokolle	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144,forc edirectio, nointr,suid`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`

In der folgenden Tabelle sind die HP-UX-NFS-Mount-Optionen für RAC aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac,suid`
Kontrolldateien Datendateien Wiederherstellungsprotokolle	`rw, bg,hard, [vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio,suid`
CRS/Abstimmung	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac, forcedirectio,suid`
Dediziert `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Freigegeben `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,suid`

Der Hauptunterschied zwischen Single-Instance- und RAC-Mount-Optionen ist das Hinzufügen von `noac` Und `forcedirectio` Zu den Mount-Optionen. Durch diese Ergänzung wird das Caching des Host-Betriebssystems deaktiviert, wodurch alle Instanzen im RAC Cluster eine konsistente Ansicht des Status der

Daten haben. Obwohl Sie den verwenden `init.ora` Parameter `filesystemio_options=setall` Hat die gleiche Wirkung wie die Deaktivierung des Host-Caching, ist es weiterhin erforderlich, zu verwenden `noac` Und `forcedirectio`.

Der Grund `noac` Ist für die gemeinsame Nutzung erforderlich `ORACLE_HOME` Die Bereitstellung soll die Konsistenz von Dateien wie Oracle-Passwortdateien und SPfiles erleichtern. Wenn jede Instanz in einem RAC-Cluster über einen dedizierten verfügt `ORACLE_HOME`, Dieser Parameter ist nicht erforderlich.

HP-UX VxFS-Mount-Optionen

Verwenden Sie die folgenden Mount-Optionen für Dateisysteme, auf denen Oracle-Binärdateien gehostet werden:

```
delaylog,nodatainlog
```

Verwenden Sie die folgenden Mount-Optionen für Dateisysteme mit Datendateien, Wiederherstellungsprotokollen, Archivprotokollen und Steuerdateien, bei denen die Version von HP-UX keine gleichzeitigen I/O unterstützt:

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

Wenn gleichzeitige I/O-Vorgänge unterstützt werden (VxFS 5.0.1 und höher oder mit der ServiceGuard Storage Management Suite), verwenden Sie diese Mount-Optionen für Dateisysteme, die Datendateien, Wiederherstellungsprotokolle, Archivprotokolle und Steuerdateien enthalten:

```
delaylog,cio
```

> Der Parameter `db_file_multiblock_read_count` Insbesondere in VxFS-Umgebungen kritisch ist. Oracle empfiehlt, dass dieser Parameter in Oracle 10g R1 und höher nicht festgelegt wird, sofern nicht ausdrücklich anders angegeben. Der Standardwert bei einer Oracle 8 KB Blockgröße ist 128. Wenn der Wert dieses Parameters auf 16 oder weniger erzwungen wird, entfernen Sie den `convosync=direct` Mount-Option, da dadurch die sequenzielle I/O-Performance beeinträchtigt werden kann. Dieser Schritt schädigt andere Aspekte der Leistung und sollte nur erfolgen, wenn der Wert von `db_file_multiblock_read_count` Muss vom Standardwert geändert werden.

Linux

Konfigurationsthemen für das Linux-Betriebssystem.

Linux NFSv3 TCP-Slot-Tabellen

TCP-Slot-Tabellen sind das NFSv3 Äquivalent zur Warteschlangentiefe des Host Bus Adapters (HBA). Diese Tabellen steuern die Anzahl der NFS-Vorgänge, die zu einem beliebigen Zeitpunkt ausstehen können. Der Standardwert ist normalerweise 16, was für eine optimale Performance viel zu niedrig ist. Das entgegengesetzte Problem tritt auf neueren Linux-Kerneln auf, die automatisch die Begrenzung der TCP-Slot-Tabelle auf ein Niveau erhöhen können, das den NFS-Server mit Anforderungen sättigt.

Um eine optimale Performance zu erzielen und Performance-Probleme zu vermeiden, passen Sie die Kernel-Parameter an, die die TCP-Slot-Tabellen steuern.

Führen Sie die aus `sysctl -a | grep tcp.*.slot_table` Und beobachten Sie die folgenden Parameter:

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

Alle Linux-Systeme sollten enthalten `sunrpc.tcp_slot_table_entries`, Aber nur einige enthalten `sunrpc.tcp_max_slot_table_entries`. Beide sollten auf 128 gesetzt werden.

> Wenn diese Parameter nicht eingestellt werden, kann dies erhebliche Auswirkungen auf die Leistung haben. In einigen Fällen ist die Performance eingeschränkt, da das linux-Betriebssystem nicht genügend I/O ausgibt In anderen Fällen erhöht sich die I/O-Latenz, wenn das linux Betriebssystem versucht, mehr I/O-Vorgänge auszustellen, als gewartet werden kann.

Mount-Optionen für Linux NFS

In der folgenden Tabelle sind die Linux-NFS-Mount-Optionen für eine einzelne Instanz aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Kontrolldateien Datendateien Wiederherstellungsprotokolle	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`
ORACLE_HOME	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

In der folgenden Tabelle sind die Linux-NFS-Mount-Optionen für RAC aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, actimeo=0`
Kontrolldateien Datendateien Wiederherstellungsprotokolle	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`
CRS/Abstimmung	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,actimeo=0`
Dediziert `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`

Dateityp	Mount-Optionen
Freigegeben `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`

Der Hauptunterschied zwischen Single-Instance- und RAC-Mount-Optionen ist das Hinzufügen von `actimeo=0` Zu den Mount-Optionen. Durch diese Ergänzung wird das Caching des Host-Betriebssystems deaktiviert, wodurch alle Instanzen im RAC Cluster eine konsistente Ansicht des Status der Daten haben. Obwohl Sie den verwenden `init.ora` Parameter `filesystemio_options=setall` Hat die gleiche Wirkung wie die Deaktivierung des Host-Caching, ist es weiterhin erforderlich, zu verwenden `actimeo=0`.

Der Grund `actimeo=0` Ist für die gemeinsame Nutzung erforderlich `ORACLE_HOME` Die Bereitstellung soll die Konsistenz von Dateien wie den Oracle-Passwortdateien und den SPfiles erleichtern. Wenn jede Instanz in einem RAC-Cluster über einen dedizierten verfügt `ORACLE_HOME`, Dann ist dieser Parameter nicht erforderlich.

Im Allgemeinen sollten nicht-Datenbankdateien mit denselben Optionen gemountet werden, die für Datendateien mit einer einzigen Instanz verwendet werden, obwohl bestimmte Applikationen unterschiedliche Anforderungen haben können. Vermeiden Sie die Montageoptionen `noac` Und `actimeo=0` Wenn möglich, da diese Optionen das Vorauslesen und Puffern auf Dateisystemebene deaktivieren. Dies kann zu schwerwiegenden Leistungsproblemen bei Prozessen wie Extraktion, Übersetzung und Laden führen.

ACCESS und GETATTR

Einige Kunden bemerken, dass ein extrem hohes Maß an anderen IOPS wie ACCESS und GETATTR ihre Workloads dominieren kann. In Extremfällen können Vorgänge wie Lese- und Schreibvorgänge bis zu 10 % des Gesamtbetrags ausmachen. Dies ist ein normales Verhalten bei jeder Datenbank, die die Verwendung einschließt `actimeo=0` Und/oder `noac` Unter Linux, weil diese Optionen dazu führen, dass das Linux-Betriebssystem ständig Datei-Metadaten aus dem Speichersystem neu lädt. Vorgänge wie ACCESS und GETATTR sind Vorgänge mit geringen Auswirkungen, die aus dem ONTAP Cache in einer Datenbankumgebung bedient werden. Sie sollten nicht als echte IOPS-Werte wie Lese- und Schreibvorgänge betrachtet werden, die einen echten Bedarf an Storage-Systemen verursachen. Diese anderen IOPS erzeugen jedoch eine gewisse Last, insbesondere in RAC-Umgebungen. Um diesem Problem zu begegnen, aktivieren Sie DNFS, wodurch der Puffercache des Betriebssystems umgangen wird und unnötige Metadatenvorgänge vermieden werden.

Linux Direct NFS

Eine zusätzliche Mount-Option, genannt `nosharecache`, Ist erforderlich, wenn (a) DNFS aktiviert ist und (b) ein Quell-Volume mehr als einmal auf einem einzelnen Server (c) mit einem verschachtelten NFS-Mount gemountet wird. Diese Konfiguration ist hauptsächlich in Umgebungen zu finden, die SAP-Anwendungen unterstützen. Beispielsweise könnte ein einzelnes Volume auf einem NetApp System über ein Verzeichnis verfügen, das sich in befindet `/vol/oracle/base` Und eine Sekunde bei `/vol/oracle/home`. Wenn `/vol/oracle/base` Ist bei montiert `/oracle` Und `/vol/oracle/home` Ist bei montiert `/oracle/home`Das Ergebnis sind verschachtelte NFS-Mounts, die von der gleichen Quelle stammen.

Das Betriebssystem kann erkennen, dass `/oracle` und `/oracle/home` befinden sich auf demselben Volume, das gleiche Quelldateisystem ist. Das Betriebssystem verwendet dann dasselbe Geräte-Handle für den Zugriff auf die Daten. Dadurch wird die Verwendung von OS Caching und bestimmten anderen Vorgängen verbessert, es beeinträchtigt jedoch DNFS. Wenn DNFS auf eine Datei zugreifen muss, wie z.B. `spfile` , ON `/oracle/home`, könnte es fälschlicherweise versuchen, den falschen Pfad zu den Daten zu verwenden. Das Ergebnis ist ein I/O-Vorgang, der fehlgeschlagen ist. Fügen Sie in diesen Konfigurationen die Mount-Option zu jedem NFS-Dateisystem hinzu `nosharecache`, das ein Quell-Volume mit einem anderen NFS-Dateisystem auf diesem Host gemeinsam nutzt. Dies zwingt das Linux-Betriebssystem, einem unabhängigen Device

Handle für dieses Dateisystem zuzuweisen.

Linux Direct NFS und Oracle RAC

Die Verwendung von DNFS bietet besondere Leistungsvorteile für Oracle RAC auf dem Linux-Betriebssystem, da Linux keine Methode zur Erzwang direkter I/O-Vorgänge bietet, die für die Kohärenz über die Knoten mit RAC erforderlich ist. Als Workaround benötigt Linux die Verwendung von `actimeo=0` Mount-Option, die dazu führt, dass Dateidaten sofort aus dem OS-Cache ablaufen. Diese Option zwingt den Linux NFS Client wiederum, Attributdaten ständig neu zu lesen, was die Latenz schädigt und die Belastung des Storage Controllers erhöht.

Durch die Aktivierung von DNFS wird der Host-NFS-Client umgangen und dieser Schaden wird vermieden. Mehrere Kunden haben bei der Aktivierung von DNFS deutliche Performance-Steigerungen bei RAC Clustern und deutliche geringere ONTAP-Lasten (insbesondere im Hinblick auf andere IOPS) gemeldet.

Linux Direct NFS- und oranfstab-Datei

Bei der Verwendung von DNFS unter Linux mit der Multipathing-Option müssen mehrere Subnetze verwendet werden. Auf anderen Betriebssystemen können mehrere DNFS-Kanäle mithilfe des eingerichtet werden `LOCAL` Und `DONTROUTE` Optionen zum Konfigurieren mehrerer DNFS-Kanäle in einem einzigen Subnetz. Dies funktioniert jedoch unter Linux nicht richtig, und es können unerwartete Leistungsprobleme auftreten. Bei Linux muss sich jeder für den DNFS-Verkehr verwendete NIC in einem anderen Subnetz befinden.

I/O-Planer

Der Linux-Kernel ermöglicht eine Steuerung auf niedriger Ebene über die Art und Weise, wie I/O-Vorgänge zum Blockieren von Geräten geplant werden. Die Standardeinstellungen auf verschiedenen Linux-Distribution variieren erheblich. Tests zeigen, dass Deadline in der Regel die besten Ergebnisse bietet, aber gelegentlich NOOP war etwas besser. Der Unterschied in der Performance ist minimal, aber testen Sie beide Optionen, wenn es erforderlich ist, um die maximal mögliche Performance aus einer Datenbankkonfiguration zu extrahieren. CFQ ist in vielen Konfigurationen der Standard und hat bei Datenbank-Workloads erhebliche Performance-Probleme gezeigt.

Anweisungen zur Konfiguration des I/O-Planers finden Sie in der entsprechenden Dokumentation des Linux-Anbieters.

Multipathing

Einige Kunden sind während der Netzwerkunterbrechung auf Abstürze gestoßen, weil der Multipath-Daemon auf ihrem System nicht ausgeführt wurde. Bei aktuellen Versionen von Linux können der Installationsprozess des Betriebssystems und des Multipathing-Daemons diese Betriebssysteme für dieses Problem anfällig machen. Die Pakete sind ordnungsgemäß installiert, aber nach einem Neustart nicht für den automatischen Start konfiguriert.

Die Standardeinstellung für den Multipath-Daemon unter RHEL5.5 kann beispielsweise wie folgt angezeigt werden:

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

Dies kann mit den folgenden Befehlen korrigiert werden:

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

ASM Spiegelung

ASM-Spiegelung erfordert möglicherweise Änderungen an den Linux Multipath-Einstellungen, damit ASM ein Problem erkennen und zu einer alternativen Ausfallgruppe wechseln kann. Die meisten ASM-Konfigurationen auf ONTAP verwenden externe Redundanz. Das bedeutet, dass Datensicherung durch das externe Array bereitgestellt wird und ASM keine Daten spiegelt. Einige Standorte verwenden ASM mit normaler Redundanz, um normalerweise zwei-Wege-Spiegelung über verschiedene Standorte hinweg bereitzustellen.

Die Linux-Einstellungen, die im angezeigt werden "NetApp Host Utilities-Dokumentation" Schließen Sie Multipath-Parameter ein, die zu unbestimmter I/O-Warteschlange führen Dies bedeutet, dass ein I/O auf einem LUN-Gerät ohne aktive Pfade so lange wartet, wie es für den I/O-Abschluss erforderlich ist. Dies ist in der Regel wünschenswert, da Linux-Hosts so lange warten, bis die Änderungen des SAN-Pfads abgeschlossen sind, FC-Switches neu gestartet werden oder ein Storage-System einen Failover abschließt.

Dieses unbegrenzte Warteschlangenverhalten verursacht ein Problem mit der ASM-Spiegelung, da ASM einen I/O-Fehler empfangen muss, damit er I/O auf einer alternativen LUN erneut versuchen kann.

Legen Sie die folgenden Parameter in Linux fest `multipath.conf` Datei für ASM-LUNs, die mit ASM-Spiegelung verwendet werden:

```
polling_interval 5
no_path_retry 24
```

Mit diesen Einstellungen wird ein Timeout von 120 Sekunden für ASM-Geräte erstellt. Das Timeout wird als berechnet `polling_interval` * `no_path_retry` Sekunden lang. Der genaue Wert muss unter Umständen angepasst werden, aber ein Timeout von 120 Sekunden sollte für die meisten Anwendungen ausreichen. Insbesondere sollten in 120 Sekunden eine Controller-Übernahme oder -Rückgabe möglich sein, ohne dass ein I/O-Fehler auftritt, der dazu führen würde, dass die Fehlergruppe offline geschaltet wird.

A niedriger `no_path_retry` Value kann die für ASM erforderliche Zeit zum Wechsel zu einer alternativen Ausfallgruppe verkürzen. Dies erhöht jedoch auch das Risiko eines unerwünschten Failovers während Wartungsaktivitäten wie beispielsweise einem Controller-Takeover. Das Risiko kann durch eine sorgfältige Überwachung des ASM-Spiegelungsstatus verringert werden. Wenn ein unerwünschtes Failover auftritt, können die Spiegelungen schnell neu synchronisiert werden, wenn die Resynchronisierung relativ schnell durchgeführt wird. Weitere Informationen finden Sie in der Oracle-Dokumentation zu ASM Fast Mirror Resync für die verwendete Version der Oracle-Software.

Mount-Optionen für Linux xfs, ext3 und ext4

NetApp empfiehlt die Verwendung der Standard-Mount-Optionen.

ASMLib/AFD (ASM-Filtertreiber)

Spezifische Konfigurationsthemen für das Linux-Betriebssystem unter Verwendung von AFD und ASMlib

ASMlib-Blockgrößen

ASMlib ist eine optionale ASM-Managementbibliothek und zugehörige Dienstprogramme. Sein primärer Wert ist die Fähigkeit, eine LUN oder eine NFS-basierte Datei als ASM-Ressource mit einem für den Benutzer lesbaren Label zu stempeln.

Aktuelle Versionen von ASMlib erkennen einen LUN-Parameter namens Logical Blocks per Physical Block Exponent (LBPPBE). Dieser Wert wurde erst vor kurzem vom ONTAP SCSI-Ziel gemeldet. Es gibt jetzt einen Wert zurück, der angibt, dass eine 4-KB-Blockgröße bevorzugt wird. Dies ist keine Definition der Blockgröße, aber es ist ein Hinweis für jede Anwendung, die LBPPBE verwendet, dass I/OS einer bestimmten Größe effizienter verarbeitet werden könnten. ASMlib interpretiert LBPPBE jedoch als Blockgröße und stempelt den ASM-Header dauerhaft, wenn das ASM-Gerät erstellt wird.

Dieser Prozess kann auf verschiedene Weise Probleme mit Upgrades und Migrationen verursachen, die auf die Unfähigkeit basieren, ASMlib-Geräte mit unterschiedlichen Blockgrößen in derselben ASM-Diskgruppe zu mischen.

Beispielsweise haben ältere Arrays im Allgemeinen einen LBPPBE-Wert von 0 gemeldet oder diesen Wert überhaupt nicht gemeldet. ASMlib interpretiert dies als 512-Byte-Blockgröße. Neuere Arrays weisen daher eine 4-KB-Blockgröße auf. Es ist nicht möglich, sowohl 512-Byte- als auch 4-KB-Geräte in derselben ASM-Diskgruppe zu mischen. Dies würde verhindern, dass ein Benutzer die Größe der ASM-Diskgruppe mit LUNs aus zwei Arrays vergrößert oder ASM als Migrationstool nutzt. In anderen Fällen erlaubt RMAN möglicherweise nicht das Kopieren von Dateien zwischen einer ASM-Diskgruppe mit einer Blockgröße von 512 Byte und einer ASM-Diskgruppe mit einer Blockgröße von 4 KB.

Die bevorzugte Lösung ist das Patchen von ASMlib. Die Oracle-Fehler-ID lautet 13999609, und der Patch ist in oracleasm-Support-2.1.8-1 und höher vorhanden. Mit diesem Patch kann der Benutzer den Parameter festlegen `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Bis `true` Im `/etc/sysconfig/oracleasm` Konfigurationsdatei Dadurch wird die Verwendung des LBPPBE-Parameters durch ASMlib blockiert, was bedeutet, dass LUNs auf dem neuen Array nun als 512-Byte-Blockgeräte erkannt werden.

> Die Option ändert nicht die Blockgröße von LUNs, die zuvor von ASMlib gestempelt wurden. Wenn beispielsweise eine ASM-Datenträgergruppe mit 512-Byte-Blöcken zu einem neuen Speichersystem migriert werden muss, das einen 4-KB-Block meldet, dann ist die Option `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Muss festgelegt werden, bevor die neuen LUNs mit ASMlib gestempelt werden. Wenn Geräte bereits durch Oracleasm gestempelt wurden, müssen sie neu formatiert werden, bevor sie mit einer neuen Blockgröße neu aufgestempelt werden. Zuerst, dekonstruieren Sie das Gerät mit `oracleasm deletedisk`, Und löschen Sie dann die ersten 1GB des Geräts mit `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Wenn das Gerät zuvor partitioniert worden war, verwenden Sie schließlich die `kpartx` Befehl, um veraltete Partitionen zu entfernen oder einfach das Betriebssystem neu zu starten.

Wenn ASMlib nicht gepatcht werden kann, kann ASMlib aus der Konfiguration entfernt werden. Diese Änderung führt zu Unterbrechungen und erfordert das Entstempeln von ASM-Festplatten und die Sicherstellung, dass die `asm_diskstring` Parameter ist korrekt eingestellt. Diese Änderung erfordert jedoch nicht die Migration der Daten.

Blockgrößen des ASM-Filterlaufwerks (AFD)

AFD ist eine optionale ASM-Managementbibliothek, die zum Ersatz für ASMlib wird. Aus Sicht des Speichers ist es ASMlib sehr ähnlich, aber es enthält zusätzliche Funktionen wie die Möglichkeit, nicht-Oracle-I/O zu blockieren, um die Wahrscheinlichkeit von Benutzer- oder Anwendungsfehlern zu verringern, die Daten beschädigen könnten.

Blockgrößen des Geräts

Wie ASMlib liest auch AFD den LUN-Parameter Logical Blocks per Physical Block Exponent (LBPPBE) und verwendet standardmäßig die physische Blockgröße, nicht die logische Blockgröße.

Dies kann zu einem Problem führen, wenn AFD zu einer bestehenden Konfiguration hinzugefügt wird, bei der die ASM-Geräte bereits als 512-Byte-Blockgeräte formatiert sind. Der AFD-Treiber erkennt die LUN als 4K-Gerät und die Diskrepanz zwischen dem ASM-Label und dem physischen Gerät würde den Zugriff verhindern. Ebenso wären Migrationen betroffen, da es nicht möglich ist, sowohl 512-Byte- als auch 4-KB-Geräte in derselben ASM-Diskgruppe zu mischen. Dies würde verhindern, dass ein Benutzer die Größe der ASM-Diskgruppe mit LUNs aus zwei Arrays vergrößert oder ASM als Migrationstool nutzt. In anderen Fällen erlaubt RMAN möglicherweise nicht das Kopieren von Dateien zwischen einer ASM-Diskgruppe mit einer Blockgröße von 512 Byte und einer ASM-Diskgruppe mit einer Blockgröße von 4 KB.

Die Lösung ist einfach: AFD enthält einen Parameter, mit dem gesteuert werden kann, ob logische oder physische Blockgrößen verwendet werden. Dies ist ein globaler Parameter, der alle Geräte im System betrifft. Um die Verwendung der logischen Blockgröße durch AFD zu erzwingen, legen Sie fest `options oracleafd oracleafd_use_logical_block_size=1` Im `/etc/modprobe.d/oracleafd.conf` Datei:

Multipath-Übertragungsgrößen

Durch die jüngsten linux-Kernel-Änderungen werden E/A-Größenbeschränkungen an Multipath-Geräte durchgesetzt, und AFD hält diese Einschränkungen nicht ein. Die I/OS werden dann abgelehnt, was dazu führt, dass der LUN-Pfad offline geschaltet wird. Dies führt dazu, dass Oracle Grid nicht installiert, ASM konfiguriert oder eine Datenbank nicht erstellt werden kann.

Die Lösung besteht darin, die maximale Übertragungslänge in der Datei multipath.conf für ONTAP-LUNs manuell anzugeben:

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

Auch wenn derzeit keine Probleme vorliegen, sollte dieser Parameter eingestellt werden, wenn AFD verwendet wird, um sicherzustellen, dass ein künftiges linux-Upgrade nicht unerwartet Probleme verursacht.

Microsoft Windows

Konfigurationsthemen für Oracle Database unter Microsoft Windows mit ONTAP..

NFS

Oracle unterstützt den Einsatz von Microsoft Windows mit dem direkten NFS-Client. Diese Funktion eröffnet neue Möglichkeiten für das Management von NFS. Hierzu zählen beispielsweise die Möglichkeit, Dateien über verschiedene Umgebungen hinweg anzuzeigen, Volumes dynamisch zu skalieren und das kostengünstigere IP-Protokoll zu nutzen. Informationen zur Installation und Konfiguration einer Datenbank unter Microsoft

Windows unter Verwendung von DNFS finden Sie in der offiziellen Oracle-Dokumentation. Es gibt keine speziellen Best Practices.

San

Stellen Sie für eine optimale Komprimierungseffizienz sicher, dass das NTFS-Dateisystem eine Zuweisungseinheit mit 8 KB oder mehr verwendet. Die Verwendung einer 4-KB-Zuweisungseinheit, die im Allgemeinen die Standardeinstellung ist, wirkt sich negativ auf die Komprimierungseffizienz aus.

Solaris

Konfigurationsthemen für das Solaris-Betriebssystem.

Solaris NFS-Mount-Optionen

In der folgenden Tabelle sind die Solaris NFS-Mount-Optionen für eine einzelne Instanz aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	`rw,bg,hard,[vers=3,vers=4.1], roto=tcp, timeo=600,rsize=262144,wsize=262144`
Steuerdateien Datendateien Wiederherstellungsprotokolle	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,llock,suid`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`

Die Verwendung von `llock` Hat sich in Kundenumgebungen nachweislich für eine drastische Performance-Steigerung bewährt, da die mit dem Erwerb und Freigeben von Sperren des Storage-Systems verbundene Latenz beseitigt wurde. Verwenden Sie diese Option sorgfältig in Umgebungen, in denen zahlreiche Server für die Bereitstellung derselben Dateisysteme konfiguriert sind und Oracle für das Mounten dieser Datenbanken konfiguriert ist. Dies ist zwar eine äußerst ungewöhnliche Konfiguration, wird jedoch von wenigen Kunden verwendet. Wenn eine Instanz versehentlich ein zweites Mal gestartet wird, kann es zu Datenbeschädigungen kommen, weil Oracle die Sperrdateien auf dem fremden Server nicht erkennen kann. NFS-Sperren bieten sonst keinen Schutz, wie in NFS-Version 3 sind sie nur beratend.

Weil die `llock` Und `forcedirectio` Parameter schließen sich gegenseitig aus, es ist wichtig, dass `filesystemio_options=setall` Befindet sich im `init.ora` So speichern `directio` Verwendet wird. Ohne diesen Parameter wird Puffer-Caching des Host-Betriebssystems verwendet und die Performance kann beeinträchtigt werden.

In der folgenden Tabelle sind die Mount-Optionen für Solaris NFS RAC aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac`

Dateityp	Mount-Optionen
Kontrolldateien Datendateien Wiederherstellungsprotokolle	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio`
CRS/Abstimmung	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio`
Dediziert `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid`
Freigegeben `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,suid`

Der Hauptunterschied zwischen Single-Instance- und RAC-Mount-Optionen ist das Hinzufügen von `noac` Und `forcedirectio` Zu den Mount-Optionen. Durch diese Ergänzung wird das Caching des Host-Betriebssystems deaktiviert, wodurch alle Instanzen im RAC Cluster eine konsistente Ansicht des Status der Daten haben. Obwohl Sie den verwenden `init.ora` Parameter `filesystemio_options=setall` Hat die gleiche Wirkung wie die Deaktivierung des Host-Caching, ist es weiterhin erforderlich, zu verwenden `noac` Und `forcedirectio`.

Der Grund `actimeo=0` Ist für die gemeinsame Nutzung erforderlich `ORACLE_HOME` Die Bereitstellung soll die Konsistenz von Dateien wie Oracle-Passwortdateien und SPfiles erleichtern. Wenn jede Instanz in einem RAC-Cluster über einen dedizierten verfügt `ORACLE_HOME`, Dieser Parameter ist nicht erforderlich.

Solaris UFS-Mount-Optionen

NetApp empfiehlt nachdrücklich die Verwendung der Mount-Option für die Protokollierung, damit die Datenintegrität im Fall eines Solaris Host-Absturzes oder der Unterbrechung der FC-Konnektivität erhalten bleibt. Die Mount-Option für die Protokollierung behält außerdem die Benutzerfreundlichkeit von Snapshot Backups bei.

Solaris ZFS

Solaris ZFS muss sorgfältig installiert und konfiguriert werden, um eine optimale Leistung zu erzielen.

Mvektor

Solaris 11 beinhaltete eine Änderung bei der Verarbeitung großer I/O-Vorgänge, die zu schwerwiegenden Leistungsproblemen auf SAN-Speicher-Arrays führen können. Das Problem ist dokumentiert NetApp Tracking Fehlerbericht 630173, "Solaris 11 ZFS Leistungsregression."

Dies ist kein ONTAP-Bug. Es handelt sich um einen Solaris-Fehler, der unter Solaris Defects 7199305 und 7082975 nachverfolgt wird.

Sie können den Oracle Support konsultieren, um herauszufinden, ob Ihre Version von Solaris 11 betroffen ist, oder Sie können die Problemumgehung testen, indem Sie auf einen kleineren Wert wechseln `zfs_mvector_max_size`.

Dazu führen Sie den folgenden Befehl als root aus:

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

Wenn unerwartete Probleme durch diese Änderung auftreten, kann sie einfach rückgängig gemacht werden, indem der folgende Befehl als Root ausgeführt wird:

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

Kernel

Eine zuverlässige ZFS-Performance erfordert einen Solaris-Kernel, der gegen Probleme bei der LUN-Ausrichtung gepatcht ist. Der Fix wurde mit Patch 147440-19 in Solaris 10 und SRU 10.5 für Solaris 11 eingeführt. Verwenden Sie nur Solaris 10 und höher mit ZFS.

LUN-Konfiguration

Führen Sie zum Konfigurieren einer LUN die folgenden Schritte aus:

1. Erstellen Sie eine LUN des Typs `solaris`.
2. Installieren Sie das entsprechende Host Utility Kit (HUK), das vom angegeben wird "NetApp Interoperabilitäts-Matrix-Tool (IMT)".
3. Befolgen Sie die Anweisungen im HUK genau wie beschrieben. Die grundlegenden Schritte sind unten beschrieben, beziehen Sie sich jedoch auf "Aktuellste Dokumentation" Für das richtige Verfahren.
   a. Führen Sie die aus `host_config` Dienstprogramm zum Aktualisieren des `sd.conf/sdd.conf` Datei: Dadurch können die SCSI-Laufwerke ONTAP-LUNs korrekt erkennen.
   b. Befolgen Sie die Anweisungen des `host_config` Dienstprogramm zur Aktivierung von Multipath Input/Output (MPIO).
   c. Neustart. Dieser Schritt ist erforderlich, damit alle Änderungen im gesamten System erkannt werden.
4. Partitionieren Sie die LUNs und stellen Sie sicher, dass sie ordnungsgemäß ausgerichtet sind. Anweisungen zum direkten Testen und Bestätigen der Ausrichtung finden Sie in Anhang B „Überprüfung der WAFL-Ausrichtung“.

Zpools

Ein zpool sollte erst nach den Schritten im erstellt werden "LUN-Konfiguration" Durchgeführt werden. Wenn das Verfahren nicht korrekt durchgeführt wird, kann es durch die I/O-Ausrichtung zu einer ernsthaften Verschlechterung der Performance kommen. Eine optimale Performance auf ONTAP erfordert, dass der I/O an einer 4-KB-Grenze auf einem Laufwerk ausgerichtet ist. Die auf einem zpool erstellten Dateisysteme verwenden eine effektive Blockgröße, die über einen Parameter mit dem Namen gesteuert wird `ashift`, Die durch Ausführen des Befehls angezeigt werden kann `zdb -C`.

Der Wert von `ashift` Der Standardwert ist 9. Dies bedeutet 2^9 oder 512 Byte. Für eine optimale Leistung, die `ashift` Wert muss 12 (2^12=4K) sein. Dieser Wert wird zum Zeitpunkt der Erstellung des zpool gesetzt und kann nicht geändert werden, was bedeutet, dass Daten in zpools mit `ashift` Andere als 12 sollten durch Kopieren der Daten in einen neu erstellten zpool migriert werden.

Überprüfen Sie nach dem Erstellen eines zpool den Wert von `ashift` Bevor Sie fortfahren. Wenn der Wert nicht 12 lautet, wurden die LUNs nicht richtig erkannt. Zerstören Sie den zpool, überprüfen Sie, ob alle Schritte in der entsprechenden Host Utilities Dokumentation korrekt ausgeführt wurden, und erstellen Sie den zpool

neu.

Zpools und Solaris LDOMs

Solaris LDOMs stellen eine zusätzliche Anforderung dar, um sicherzustellen, dass die I/O-Ausrichtung korrekt ist. Obwohl eine LUN möglicherweise ordnungsgemäß als 4K-Gerät erkannt wird, erbt ein virtuelles vdsk-Gerät auf einem LDOM die Konfiguration nicht von der I/O-Domäne. Die vdsk auf Basis dieser LUN wird standardmäßig auf einen 512-Byte-Block zurückgesetzt.

Eine zusätzliche Konfigurationsdatei ist erforderlich. Zunächst müssen die einzelnen LDOMs für Oracle Bug 15824910 gepatcht werden, um die zusätzlichen Konfigurationsoptionen zu aktivieren. Dieser Patch wurde in alle derzeit verwendeten Versionen von Solaris portiert. Sobald das LDOM gepatcht ist, kann es wie folgt konfiguriert werden:

1. Identifizieren Sie die LUN oder LUNs, die in dem neuen zpool verwendet werden sollen. In diesem Beispiel handelt es sich um das c2d1-Gerät.

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

2. Rufen Sie die vdc-Instanz der Geräte ab, die für einen ZFS-Pool verwendet werden sollen:

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

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

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

Dies bedeutet, dass Geräteinstanz 1 eine Blockgröße von 4096 zugewiesen wird.

Nehmen wir als weiteres Beispiel an, dass die vdsk-Instanzen 1 bis 6 für eine 4-KB-Blockgröße und konfiguriert sein müssen `/etc/path_to_inst` Lautet wie folgt:

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

4. Das Finale `vdc.conf` Die Datei sollte Folgendes enthalten:

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

Achtung
Das LDOM muss neu gestartet werden, nachdem vdc.conf konfiguriert und vdsk erstellt wurde. Dieser Schritt kann nicht vermieden werden. Die Änderung der Blockgröße wird nur nach einem Neustart wirksam. Fahren Sie mit der Konfiguration von zpool fort und stellen Sie sicher, dass der Ashift wie zuvor beschrieben richtig auf 12 eingestellt ist.

ZFS-Absichtsprotokoll (ZIL)

Im Allgemeinen gibt es keinen Grund, das ZFS Intent Log (ZIL) auf einem anderen Gerät zu finden. Das Protokoll kann Speicherplatz mit dem Hauptpool teilen. Die primäre Verwendung eines separaten ZIL ist, wenn physische Laufwerke verwendet werden, denen die Schreib-Cache-Funktionen in modernen Speicher-Arrays fehlen.

Logbias

Stellen Sie die ein `logbias` Parameter auf ZFS-Dateisystemen, auf denen Oracle-Daten gehostet werden.

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

Die Verwendung dieses Parameters verringert die Gesamtschreibebenen. Unter den Standardeinstellungen werden geschriebene Daten zuerst an das ZIL und dann an den Hauptspeicherpool übertragen. Dieser Ansatz eignet sich für eine Konfiguration mit einer einfachen Laufwerkskonfiguration, die ein SSD-basiertes ZIL-Gerät und rotierende Medien für den Hauptspeicherpool umfasst. Dies liegt daran, dass eine Übertragung in einer einzelnen I/O-Transaktion auf den Medien mit der niedrigsten verfügbaren Latenz ausgeführt werden kann.

Bei Verwendung eines modernen Storage Array mit eigener Caching-Funktion ist dieser Ansatz in der Regel nicht erforderlich. In seltenen Fällen ist es wünschenswert, einen Schreibvorgang mit einer einzigen Transaktion in das Protokoll übertragen zu können, z. B. bei einem Workload, der aus hochkonzentrierten, latenzempfindlichen zufälligen Schreibvorgängen besteht. Die Form der Write Amplification hat Folgen, da die protokollierten Daten schließlich in den Haupt-Storage Pool geschrieben werden, wodurch die Schreibaktivität verdoppelt wird.

Direkter I/O

Viele Applikationen, darunter auch Oracle Produkte, können den Host-Puffer-Cache umgehen, indem sie direkten I/O aktivieren Diese Strategie funktioniert bei ZFS-Dateisystemen nicht wie erwartet. Obwohl der Host-Puffer-Cache umgangen wird, speichert ZFS selbst weiterhin Daten im Cache. Dies kann zu irreführenden Ergebnissen führen, wenn Tools wie fio oder sio für Performance-Tests verwendet werden, da schwer vorherzusagen ist, ob I/O das Storage-System erreicht oder ob es lokal im BS zwischengespeichert wird. Diese Aktion macht es auch sehr schwierig, solche synthetischen Tests zu verwenden, um ZFS-Leistung mit anderen Dateisystemen zu vergleichen. In der Praxis gibt es bei echten Benutzer-Workloads kaum bis keine Unterschiede in der Filesystem-Performance.

Mehrere zpools

Snapshot-basierte Backups, Wiederherstellungen, Klone und Archivierung von ZFS-basierten Daten müssen auf der Ebene von zpool durchgeführt werden und erfordern in der Regel mehrere zpools. Ein zpool ist analog zu einer LVM-Plattengruppe und sollte mit denselben Regeln konfiguriert werden. Beispielsweise ist eine Datenbank wahrscheinlich am besten mit den Datendateien in ausgelegt `zpool1` Und die Archivprotokolle, Kontrolldateien und Wiederherstellungsprotokolle befinden sich auf `zpool2`. Dieser Ansatz ermöglicht ein Standard-Hot Backup, bei dem sich die Datenbank im Hot Backup-Modus befindet, gefolgt von einem Snapshot von `zpool1`. Die Datenbank wird dann aus dem Hot Backup-Modus entfernt, das Protokollarchiv wird erzwungen und ein Snapshot von `zpool2` Wird erstellt. Ein Wiederherstellungsvorgang erfordert das Abhängen der zfs-Dateisysteme und den vollständigen Offlining des zpool nach einer SnapRestore-Wiederherstellung. Der zpool kann dann wieder online gebracht werden und die Datenbank wiederhergestellt werden.

Filesystemio_options

Der Oracle-Parameter `filesystemio_options` Funktioniert anders mit ZFS. Wenn `setall` Oder `directio` Wird verwendet, Schreibvorgänge sind synchron und umgehen den BS-Puffer-Cache, aber Lesevorgänge werden von ZFS gepuffert. Diese Aktion führt zu Schwierigkeiten bei der Performance-Analyse, da I/O manchmal vom ZFS-Cache abgefangen und gewartet wird. Dadurch werden die Speicherlatenz und der gesamte I/O geringer als möglicherweise angezeigt.

Hostkonfiguration mit ASA r2-Systemen

AIX

Konfigurationsthemen für Oracle-Datenbanken auf IBM AIX mit ASA r2 ONTAP.

AIX wird mit NetApp ASA r2 für das Hosting von Oracle-Datenbanken unterstützt, vorausgesetzt:

- Sie haben Oracle für gleichzeitige E/A-Operationen korrekt konfiguriert.
- Sie verwenden unterstützte SAN-Protokolle (FC/iSCSI/NVMe).
- Sie verwenden ONTAP 9.16.x oder höher auf ASA r2.

Gleichzeitige I/O-Vorgänge

Um auf IBM AIX mit ASA r2 eine optimale Leistung zu erzielen, ist die Verwendung von paralleler E/A erforderlich. Ohne gleichzeitige E/A-Operationen sind Leistungseinschränkungen wahrscheinlich, da AIX serialisierte, atomare E/A-Operationen durchführt, was einen erheblichen Mehraufwand verursacht.

Ursprünglich empfahl NetApp die Verwendung von `cio` Die Mount-Option erzwingt gleichzeitige E/A-Operationen im Dateisystem, aber dieses Verfahren hatte Nachteile und ist nicht mehr erforderlich. Seit der Einführung von AIX 5.2 und Oracle 10gR1 kann Oracle auf AIX einzelne Dateien für gleichzeitige E/A öffnen, anstatt gleichzeitige E/A auf dem gesamten Dateisystem zu erzwingen.

Die beste Methode für die Aktivierung gleichzeitiger I/O ist, die festzulegen `init.ora` Parameter `filesystemio_options` Bis `setall`. Auf diese Weise kann Oracle spezifische Dateien zur Verwendung mit gleichzeitigen I/O-Vorgängen öffnen

Die Verwendung von cio als Mount-Option erzwingt die Nutzung von gleichzeitiger E/A, was negative Folgen haben kann. Beispielsweise deaktiviert das Erzwingen von gleichzeitigen E/A-Vorgängen das Vorlesen auf Dateisystemen, was die Leistung bei E/A-Vorgängen außerhalb der Oracle-Datenbanksoftware beeinträchtigen kann, etwa beim Kopieren von Dateien und beim Durchführen von Bandsicherungen. Darüber hinaus sind Produkte wie Oracle GoldenGate und SAP BR*Tools mit bestimmten Versionen von Oracle nicht kompatibel, wenn die Option „cio mount“ verwendet wird.

NetApp empfiehlt Folgendes:

- Verwenden Sie das nicht `cio` Mount-Option auf Filesystem-Ebene. Aktivieren Sie stattdessen Concurrent I/O über `filesystemio_options=setall`.
- Verwenden Sie ausschließlich `cio` Mount-Option, falls die Einstellung nicht möglich ist `filesystemio_options=setall`Die

Da ASA r2 NAS nicht unterstützt, müssen alle Oracle-Bereitstellungen auf AIX Blockprotokolle verwenden.

Mount-Optionen für AIX jfs/jfs2

In der folgenden Tabelle sind die AIX jfs/jfs2-Mount-Optionen aufgeführt.

Dateityp	Mount-Optionen
AdR-Startseite	Standardwerte
Kontrolldateien	Standardwerte
Datendateien	Standardwerte
Redo-Protokolle	Standardwerte
ORACLE_HOME	Standardwerte

Vor der Verwendung von AIX `hdisk` Geräte in jeder Umgebung, einschließlich Datenbanken, überprüfen den Parameter `queue_depth`Die Dieser Parameter gibt nicht die HBA-Warteschlangenlänge an, sondern bezieht sich auf die SCSI-Warteschlangenlänge des einzelnen Geräts. `hdisk device`Die Je nachdem, wie die ASA r2 LUNs konfiguriert sind, ist der Wert für `queue_depth` könnte für eine gute Leistung zu niedrig sein. Tests haben gezeigt, dass der optimale Wert bei 64 liegt.

HP-UX ERHÄLTLICH

Konfigurationsthemen für Oracle-Datenbanken auf HP-UX mit ASA r2 ONTAP.

HP-UX wird mit NetApp ASA r2 für das Hosting von Oracle-Datenbanken unterstützt, vorausgesetzt:

- Die ONTAP Version muss 9.16.x oder höher sein.
- Verwenden Sie SAN-Protokolle (FC/iSCSI/NVMe). NAS wird auf ASA r2 nicht unterstützt.
- Wenden Sie die für HP-UX spezifischen Best Practices für die Montage und I/O-Optimierung an.

HP-UX VxFS-Mount-Optionen

Verwenden Sie die folgenden Mount-Optionen für Dateisysteme, auf denen Oracle-Binärdateien gehostet werden:

```
delaylog,nodatainlog
```

Verwenden Sie die folgenden Mount-Optionen für Dateisysteme mit Datendateien, Wiederherstellungsprotokollen, Archivprotokollen und Steuerdateien, bei denen die Version von HP-UX keine gleichzeitigen I/O unterstützt:

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

Wenn gleichzeitige I/O-Vorgänge unterstützt werden (VxFS 5.0.1 und höher oder mit der ServiceGuard Storage Management Suite), verwenden Sie diese Mount-Optionen für Dateisysteme, die Datendateien, Wiederherstellungsprotokolle, Archivprotokolle und Steuerdateien enthalten:

```
delaylog,cio
```

Der Parameter `db_file_multiblock_read_count` Insbesondere in VxFS-Umgebungen kritisch ist. Oracle empfiehlt, dass dieser Parameter in Oracle 10g R1 und höher nicht festgelegt wird, sofern nicht ausdrücklich anders angegeben. Der Standardwert bei einer Oracle 8 KB Blockgröße ist 128. Wenn der Wert dieses Parameters auf 16 oder weniger erzwungen wird, entfernen Sie den `convosync=direct` Mount-Option, da dadurch die sequenzielle I/O-Performance beeinträchtigt werden kann. Dieser Schritt schädigt andere Aspekte der Leistung und sollte nur erfolgen, wenn der Wert von `db_file_multiblock_read_count` Muss vom Standardwert geändert werden.

Linux

Konfigurationsthemen speziell für das Linux-Betriebssystem mit ASA r2 ONTAP.

Linux (Oracle Linux, RHEL, SUSE) wird mit ASA r2 für Oracle-Datenbanken unterstützt. Verwenden Sie SAN-Protokolle, konfigurieren Sie Multipathing korrekt und wenden Sie die Best Practices von Oracle für ASM und I/O-Tuning an.

I/O-Planer

Der Linux-Kernel ermöglicht eine Steuerung auf niedriger Ebene über die Art und Weise, wie I/O-Vorgänge zum Blockieren von Geräten geplant werden. Die Standardeinstellungen auf verschiedenen Linux-Distribution variieren erheblich. Tests zeigen, dass Deadline in der Regel die besten Ergebnisse bietet, aber gelegentlich NOOP war etwas besser. Der Unterschied in der Performance ist minimal, aber testen Sie beide Optionen, wenn es erforderlich ist, um die maximal mögliche Performance aus einer Datenbankkonfiguration zu extrahieren. CFQ ist in vielen Konfigurationen der Standard und hat bei Datenbank-Workloads erhebliche Performance-Probleme gezeigt.

Anweisungen zur Konfiguration des I/O-Planers finden Sie in der entsprechenden Dokumentation des Linux-Anbieters.

Multipathing

Einige Kunden sind während der Netzwerkunterbrechung auf Abstürze gestoßen, weil der Multipath-Daemon auf ihrem System nicht ausgeführt wurde. Bei aktuellen Versionen von Linux können der Installationsprozess des Betriebssystems und des Multipathing-Daemons diese Betriebssysteme für dieses Problem anfällig machen. Die Pakete sind ordnungsgemäß installiert, aber nach einem Neustart nicht für den automatischen Start konfiguriert.

Die Standardkonfiguration für den Multipath-Daemon unter RHEL 9.7 könnte beispielsweise wie folgt aussehen:

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

Dies kann mit den folgenden Befehlen korrigiert werden:

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

Warteschlangentiefe

Um E/A-Engpässe zu vermeiden, sollte für SAN-Geräte eine geeignete Warteschlangenlänge eingestellt werden. Die Standard-Warteschlangenlänge unter Linux ist oft auf 128 eingestellt, was bei Oracle-Datenbanken zu Leistungsproblemen führen kann. Eine zu hohe Warteschlangenlänge kann zu übermäßiger E/A-Warteschlangenbildung führen, was wiederum die Latenz erhöht und den Durchsatz verringert. Wenn der Wert zu niedrig eingestellt ist, kann dies die Anzahl der ausstehenden E/A-Anforderungen begrenzen und somit die Gesamtleistung verringern. Eine Warteschlangenlänge von 64 ist oft ein guter Ausgangspunkt für Oracle-Datenbank-Workloads auf ASA r2, muss aber je nach spezifischen Workload-Charakteristika und Leistungstests angepasst werden.

ASM Spiegelung

ASM-Spiegelung erfordert möglicherweise Änderungen an den Linux Multipath-Einstellungen, damit ASM ein Problem erkennen und zu einer alternativen Ausfallgruppe wechseln kann. Die meisten ASM-Konfigurationen auf ONTAP verwenden externe Redundanz. Das bedeutet, dass Datensicherung durch das externe Array bereitgestellt wird und ASM keine Daten spiegelt. Einige Standorte verwenden ASM mit normaler Redundanz, um normalerweise zwei-Wege-Spiegelung über verschiedene Standorte hinweg bereitzustellen.

Bei ASA r2-Systemen, die Active-Active Multipathing unterstützen, sollten diese Multipath-Einstellungen angepasst werden. Da alle Pfade aktiv und lastverteilt sind, ist eine unbegrenzte Warteschlangenbildung nicht erforderlich. Stattdessen sollten Multipath-Parameter die Leistung und ein schnelles Failback priorisieren. Dieses Verhalten ist für die ASM-Spiegelung wichtig, da ASM einen E/A-Fehler empfangen muss, um die E/A auf einer alternativen LUN erneut zu versuchen. Wenn E/A-Vorgänge unbegrenzt in der Warteschlange stehen, kann ASM kein Failover auslösen.

Legen Sie die folgenden Parameter in Linux fest `multipath.conf` Datei für ASM-LUNs, die mit ASM-Spiegelung verwendet werden:

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

Mit diesen Einstellungen wird ein Timeout von 120 Sekunden für ASM-Geräte erstellt. Das Timeout wird als berechnet `polling_interval` * `no_path_retry` Sekunden lang. Der genaue Wert muss unter Umständen angepasst werden, aber ein Timeout von 120 Sekunden sollte für die meisten Anwendungen ausreichen. Insbesondere sollten in 120 Sekunden eine Controller-Übernahme oder -Rückgabe möglich sein, ohne dass ein I/O-Fehler auftritt, der dazu führen würde, dass die Fehlergruppe offline geschaltet wird.

A niedriger `no_path_retry` Value kann die für ASM erforderliche Zeit zum Wechsel zu einer alternativen Ausfallgruppe verkürzen. Dies erhöht jedoch auch das Risiko eines unerwünschten Failovers während Wartungsaktivitäten wie beispielsweise einem Controller-Takeover. Das Risiko kann durch eine sorgfältige Überwachung des ASM-Spiegelungsstatus verringert werden. Wenn ein unerwünschtes Failover auftritt, können die Spiegelungen schnell neu synchronisiert werden, wenn die Resynchronisierung relativ schnell durchgeführt wird. Weitere Informationen finden Sie in der Oracle-Dokumentation zu ASM Fast Mirror Resync für die verwendete Version der Oracle-Software.

Mount-Optionen für Linux xfs, ext3 und ext4

* NetApp empfiehlt* die Verwendung der Standard-Mount-Optionen. Achten Sie auf die korrekte Ausrichtung beim Erstellen von Dateisystemen auf LUNs.

ASMLib/AFD (ASM-Filtertreiber)

Konfigurationsthemen speziell für das Linux-Betriebssystem unter Verwendung von AFD und ASMlib mit ASA r2 ONTAP.

ASMlib-Blockgrößen

ASMlib ist eine optionale ASM-Verwaltungsbibliothek mit zugehörigen Hilfsprogrammen. Sein Hauptnutzen besteht in der Möglichkeit, eine LUN als ASM-Ressource mit einer für Menschen lesbaren Bezeichnung zu versehen.

Aktuelle Versionen von ASMlib erkennen einen LUN-Parameter namens Logical Blocks per Physical Block Exponent (LBPPBE). Dieser Wert wurde erst vor kurzem vom ONTAP SCSI-Ziel gemeldet. Es gibt jetzt einen Wert zurück, der angibt, dass eine 4-KB-Blockgröße bevorzugt wird. Dies ist keine Definition der Blockgröße, aber es ist ein Hinweis für jede Anwendung, die LBPPBE verwendet, dass I/OS einer bestimmten Größe effizienter verarbeitet werden könnten. ASMlib interpretiert LBPPBE jedoch als Blockgröße und stempelt den ASM-Header dauerhaft, wenn das ASM-Gerät erstellt wird.

Dieser Prozess kann auf verschiedene Weise Probleme mit Upgrades und Migrationen verursachen, die auf die Unfähigkeit basieren, ASMlib-Geräte mit unterschiedlichen Blockgrößen in derselben ASM-Diskgruppe zu mischen.

Beispielsweise haben ältere Arrays im Allgemeinen einen LBPPBE-Wert von 0 gemeldet oder diesen Wert überhaupt nicht gemeldet. ASMlib interpretiert dies als 512-Byte-Blockgröße. Neuere Arrays weisen daher eine 4-KB-Blockgröße auf. Es ist nicht möglich, sowohl 512-Byte- als auch 4-KB-Geräte in derselben ASM-Diskgruppe zu mischen. Dies würde verhindern, dass ein Benutzer die Größe der ASM-Diskgruppe mit LUNs aus zwei Arrays vergrößert oder ASM als Migrationstool nutzt. In anderen Fällen erlaubt RMAN möglicherweise nicht das Kopieren von Dateien zwischen einer ASM-Diskgruppe mit einer Blockgröße von 512 Byte und einer ASM-Diskgruppe mit einer Blockgröße von 4 KB.

Die bevorzugte Lösung ist das Patchen von ASMlib. Die Oracle-Fehler-ID lautet 13999609, und der Patch ist in oracleasm-Support-2.1.8-1 und höher vorhanden. Mit diesem Patch kann der Benutzer den Parameter festlegen `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Bis `true` Im `/etc/sysconfig/oracleasm` Konfigurationsdatei Dadurch wird die Verwendung des LBPPBE-Parameters durch ASMlib blockiert, was bedeutet, dass LUNs auf dem neuen Array nun als 512-Byte-Blockgeräte erkannt werden.

> Die Option ändert nicht die Blockgröße von LUNs, die zuvor von ASMlib gestempelt wurden. Wenn beispielsweise eine ASM-Datenträgergruppe mit 512-Byte-Blöcken zu einem neuen Speichersystem migriert werden muss, das einen 4-KB-Block meldet, dann ist die Option `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Muss festgelegt werden, bevor die neuen LUNs mit ASMlib gestempelt werden. Wenn Geräte bereits durch Oracleasm gestempelt wurden, müssen sie neu formatiert werden, bevor sie mit einer neuen Blockgröße neu aufgestempelt werden. Zuerst, dekonstruieren Sie das Gerät mit `oracleasm deletedisk`, Und löschen Sie dann die ersten 1GB des Geräts mit `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Wenn das Gerät zuvor partitioniert worden war, verwenden Sie schließlich die `kpartx` Befehl, um veraltete Partitionen zu entfernen oder einfach das Betriebssystem neu zu starten.

Wenn ASMlib nicht gepatcht werden kann, kann ASMlib aus der Konfiguration entfernt werden. Diese Änderung führt zu Unterbrechungen und erfordert das Entstempeln von ASM-Festplatten und die Sicherstellung, dass die `asm_diskstring` Parameter ist korrekt eingestellt. Diese Änderung erfordert jedoch nicht die Migration der Daten.

Blockgrößen des ASM-Filterlaufwerks (AFD)

AFD ist eine optionale ASM-Managementbibliothek, die zum Ersatz für ASMlib wird. Aus Sicht des Speichers ist es ASMlib sehr ähnlich, aber es enthält zusätzliche Funktionen wie die Möglichkeit, nicht-Oracle-I/O zu blockieren, um die Wahrscheinlichkeit von Benutzer- oder Anwendungsfehlern zu verringern, die Daten beschädigen könnten.

Blockgrößen des Geräts

Wie ASMlib liest auch AFD den LUN-Parameter Logical Blocks per Physical Block Exponent (LBPPBE) und verwendet standardmäßig die physische Blockgröße, nicht die logische Blockgröße.

Dies kann zu einem Problem führen, wenn AFD zu einer bestehenden Konfiguration hinzugefügt wird, bei der die ASM-Geräte bereits als 512-Byte-Blockgeräte formatiert sind. Der AFD-Treiber erkennt die LUN als 4K-Gerät und die Diskrepanz zwischen dem ASM-Label und dem physischen Gerät würde den Zugriff verhindern. Ebenso wären Migrationen betroffen, da es nicht möglich ist, sowohl 512-Byte- als auch 4-KB-Geräte in derselben ASM-Diskgruppe zu mischen. Dies würde verhindern, dass ein Benutzer die Größe der ASM-Diskgruppe mit LUNs aus zwei Arrays vergrößert oder ASM als Migrationstool nutzt. In anderen Fällen erlaubt RMAN möglicherweise nicht das Kopieren von Dateien zwischen einer ASM-Diskgruppe mit einer Blockgröße von 512 Byte und einer ASM-Diskgruppe mit einer Blockgröße von 4 KB.

Die Lösung ist einfach: AFD enthält einen Parameter, mit dem gesteuert werden kann, ob logische oder physische Blockgrößen verwendet werden. Dies ist ein globaler Parameter, der alle Geräte im System betrifft. Um die Verwendung der logischen Blockgröße durch AFD zu erzwingen, legen Sie fest `options oracleafd oracleafd_use_logical_block_size=1` Im `/etc/modprobe.d/oracleafd.conf` Datei:

Multipath-Übertragungsgrößen

Durch die jüngsten linux-Kernel-Änderungen werden E/A-Größenbeschränkungen an Multipath-Geräte durchgesetzt, und AFD hält diese Einschränkungen nicht ein. Die I/OS werden dann abgelehnt, was dazu führt, dass der LUN-Pfad offline geschaltet wird. Dies führt dazu, dass Oracle Grid nicht installiert, ASM konfiguriert oder eine Datenbank nicht erstellt werden kann.

Die Lösung besteht darin, die maximale Übertragungslänge in der Datei multipath.conf für ONTAP-LUNs manuell anzugeben:

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

Auch wenn derzeit keine Probleme vorliegen, sollte dieser Parameter eingestellt werden, wenn AFD verwendet wird, um sicherzustellen, dass ein künftiges linux-Upgrade nicht unerwartet Probleme verursacht.

Microsoft Windows

Konfigurationsthemen für Oracle-Datenbanken unter Microsoft Windows mit ASA r2 ONTAP.

San

Stellen Sie für eine optimale Komprimierungseffizienz sicher, dass das NTFS-Dateisystem eine Zuweisungseinheit mit 8 KB oder mehr verwendet. Die Verwendung einer 4-KB-Zuweisungseinheit, die im Allgemeinen die Standardeinstellung ist, wirkt sich negativ auf die Komprimierungseffizienz aus.

Solaris

Konfigurationsthemen speziell für das Solaris-Betriebssystem mit ASA r2 ONTAP.

Solaris UFS-Mount-Optionen

NetApp empfiehlt nachdrücklich die Verwendung der Mount-Option für die Protokollierung, damit die Datenintegrität im Fall eines Solaris Host-Absturzes oder der Unterbrechung der FC-Konnektivität erhalten bleibt. Die Mount-Option für die Protokollierung behält außerdem die Benutzerfreundlichkeit von Snapshot Backups bei.

Solaris ZFS

Solaris ZFS muss sorgfältig installiert und konfiguriert werden, um eine optimale Leistung zu erzielen.

Mvektor

Solaris 11 beinhaltete eine Änderung bei der Verarbeitung großer I/O-Vorgänge, die zu schwerwiegenden Leistungsproblemen auf SAN-Speicher-Arrays führen können. Das Problem ist dokumentiert NetApp Tracking Fehlerbericht 630173, "Solaris 11 ZFS Leistungsregression."

Dies ist kein ONTAP-Bug. Es handelt sich um einen Solaris-Fehler, der unter Solaris Defects 7199305 und 7082975 nachverfolgt wird.

Sie können den Oracle Support konsultieren, um herauszufinden, ob Ihre Version von Solaris 11 betroffen ist, oder Sie können die Problemumgehung testen, indem Sie auf einen kleineren Wert wechseln `zfs_mvector_max_size`.

Dazu führen Sie den folgenden Befehl als root aus:

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

Wenn unerwartete Probleme durch diese Änderung auftreten, kann sie einfach rückgängig gemacht werden, indem der folgende Befehl als Root ausgeführt wird:

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

Kernel

Eine zuverlässige ZFS-Performance erfordert einen Solaris-Kernel, der gegen Probleme bei der LUN-Ausrichtung gepatcht ist. Der Fix wurde mit Patch 147440-19 in Solaris 10 und SRU 10.5 für Solaris 11 eingeführt. Verwenden Sie nur Solaris 10 und höher mit ZFS.

LUN-Konfiguration

Führen Sie zum Konfigurieren einer LUN die folgenden Schritte aus:

1. Erstellen Sie eine LUN des Typs `solaris`.
2. Installieren Sie das entsprechende Host Utility Kit (HUK), das vom angegeben wird "NetApp Interoperabilitäts-Matrix-Tool (IMT)".

3. Befolgen Sie die Anweisungen im HUK genau wie beschrieben. Die grundlegenden Schritte sind unten beschrieben, beziehen Sie sich jedoch auf "Aktuellste Dokumentation" Für das richtige Verfahren.
   a. Führen Sie die aus `host_config` Dienstprogramm zum Aktualisieren des `sd.conf/sdd.conf` Datei: Dadurch können die SCSI-Laufwerke ONTAP-LUNs korrekt erkennen.
   b. Befolgen Sie die Anweisungen des `host_config` Dienstprogramm zur Aktivierung von Multipath Input/Output (MPIO).
   c. Neustart. Dieser Schritt ist erforderlich, damit alle Änderungen im gesamten System erkannt werden.
4. Partitionieren Sie die LUNs und stellen Sie sicher, dass sie ordnungsgemäß ausgerichtet sind. Anweisungen zum direkten Testen und Bestätigen der Ausrichtung finden Sie in Anhang B „Überprüfung der WAFL-Ausrichtung“.

Zpools

Ein zpool sollte erst nach den Schritten im erstellt werden "LUN-Konfiguration" Durchgeführt werden. Wenn das Verfahren nicht korrekt durchgeführt wird, kann es durch die I/O-Ausrichtung zu einer ernsthaften Verschlechterung der Performance kommen. Eine optimale Performance auf ONTAP erfordert, dass der I/O an einer 4-KB-Grenze auf einem Laufwerk ausgerichtet ist. Die auf einem zpool erstellten Dateisysteme verwenden eine effektive Blockgröße, die über einen Parameter mit dem Namen gesteuert wird `ashift`, Die durch Ausführen des Befehls angezeigt werden kann `zdb -C`.

Der Wert von `ashift` Der Standardwert ist 9. Dies bedeutet 2^9 oder 512 Byte. Für eine optimale Leistung, die `ashift` Wert muss 12 (2^12=4K) sein. Dieser Wert wird zum Zeitpunkt der Erstellung des zpool gesetzt und kann nicht geändert werden, was bedeutet, dass Daten in zpools mit `ashift` Andere als 12 sollten durch Kopieren der Daten in einen neu erstellten zpool migriert werden.

Überprüfen Sie nach dem Erstellen eines zpool den Wert von `ashift` Bevor Sie fortfahren. Wenn der Wert nicht 12 lautet, wurden die LUNs nicht richtig erkannt. Zerstören Sie den zpool, überprüfen Sie, ob alle Schritte in der entsprechenden Host Utilities Dokumentation korrekt ausgeführt wurden, und erstellen Sie den zpool neu.

Zpools und Solaris LDOMs

Solaris LDOMs stellen eine zusätzliche Anforderung dar, um sicherzustellen, dass die I/O-Ausrichtung korrekt ist. Obwohl eine LUN möglicherweise ordnungsgemäß als 4K-Gerät erkannt wird, erbt ein virtuelles vdsk-Gerät auf einem LDOM die Konfiguration nicht von der I/O-Domäne. Die vdsk auf Basis dieser LUN wird standardmäßig auf einen 512-Byte-Block zurückgesetzt.

Eine zusätzliche Konfigurationsdatei ist erforderlich. Zunächst müssen die einzelnen LDOMs für Oracle Bug 15824910 gepatcht werden, um die zusätzlichen Konfigurationsoptionen zu aktivieren. Dieser Patch wurde in alle derzeit verwendeten Versionen von Solaris portiert. Sobald das LDOM gepatcht ist, kann es wie folgt konfiguriert werden:

1. Identifizieren Sie die LUN oder LUNs, die in dem neuen zpool verwendet werden sollen. In diesem Beispiel handelt es sich um das c2d1-Gerät.

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

2. Rufen Sie die vdc-Instanz der Geräte ab, die für einen ZFS-Pool verwendet werden sollen:

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

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

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

Dies bedeutet, dass Geräteinstanz 1 eine Blockgröße von 4096 zugewiesen wird.

Nehmen wir als weiteres Beispiel an, dass die vdsk-Instanzen 1 bis 6 für eine 4-KB-Blockgröße und konfiguriert sein müssen `/etc/path_to_inst` Lautet wie folgt:

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

4. Das Finale `vdc.conf` Die Datei sollte Folgendes enthalten:

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

Das LDOM muss neu gestartet werden, nachdem vdc.conf konfiguriert und vdsk erstellt wurde. Dieser Schritt kann nicht vermieden werden. Die Änderung der Blockgröße wird nur nach einem Neustart wirksam. Fahren Sie mit der Konfiguration von zpool fort und stellen Sie sicher, dass der Ashift wie zuvor beschrieben richtig auf 12 eingestellt ist.

ZFS-Absichtsprotokoll (ZIL)

Im Allgemeinen gibt es keinen Grund, das ZFS Intent Log (ZIL) auf einem anderen Gerät zu finden. Das Protokoll kann Speicherplatz mit dem Hauptpool teilen. Die primäre Verwendung eines separaten ZIL ist, wenn physische Laufwerke verwendet werden, denen die Schreib-Cache-Funktionen in modernen Speicher-Arrays fehlen.

Logbias

Stellen Sie die ein `logbias` Parameter auf ZFS-Dateisystemen, auf denen Oracle-Daten gehostet werden.

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

Die Verwendung dieses Parameters verringert die Gesamtschreibebenen. Unter den Standardeinstellungen werden geschriebene Daten zuerst an das ZIL und dann an den Hauptspeicherpool übertragen. Dieser Ansatz eignet sich für eine Konfiguration mit einer einfachen Laufwerkskonfiguration, die ein SSD-basiertes ZIL-Gerät und rotierende Medien für den Hauptspeicherpool umfasst. Dies liegt daran, dass eine Übertragung in einer einzelnen I/O-Transaktion auf den Medien mit der niedrigsten verfügbaren Latenz ausgeführt werden kann.

Bei Verwendung eines modernen Storage Array mit eigener Caching-Funktion ist dieser Ansatz in der Regel nicht erforderlich. In seltenen Fällen ist es wünschenswert, einen Schreibvorgang mit einer einzigen Transaktion in das Protokoll übertragen zu können, z. B. bei einem Workload, der aus hochkonzentrierten, latenzempfindlichen zufälligen Schreibvorgängen besteht. Die Form der Write Amplification hat Folgen, da die protokollierten Daten schließlich in den Haupt-Storage Pool geschrieben werden, wodurch die Schreibaktivität verdoppelt wird.

Direkter I/O

Viele Applikationen, darunter auch Oracle Produkte, können den Host-Puffer-Cache umgehen, indem sie direkten I/O aktivieren Diese Strategie funktioniert bei ZFS-Dateisystemen nicht wie erwartet. Obwohl der Host-Puffer-Cache umgangen wird, speichert ZFS selbst weiterhin Daten im Cache. Dies kann zu irreführenden

Ergebnissen führen, wenn Tools wie fio oder sio für Performance-Tests verwendet werden, da schwer vorherzusagen ist, ob I/O das Storage-System erreicht oder ob es lokal im BS zwischengespeichert wird. Diese Aktion macht es auch sehr schwierig, solche synthetischen Tests zu verwenden, um ZFS-Leistung mit anderen Dateisystemen zu vergleichen. In der Praxis gibt es bei echten Benutzer-Workloads kaum bis keine Unterschiede in der Filesystem-Performance.

Mehrere zpools

Snapshot-basierte Backups, Wiederherstellungen, Klone und Archivierung von ZFS-basierten Daten müssen auf der Ebene von zpool durchgeführt werden und erfordern in der Regel mehrere zpools. Ein zpool ist analog zu einer LVM-Plattengruppe und sollte mit denselben Regeln konfiguriert werden. Beispielsweise ist eine Datenbank wahrscheinlich am besten mit den Datendateien in ausgelegt `zpool1` Und die Archivprotokolle, Kontrolldateien und Wiederherstellungsprotokolle befinden sich auf `zpool2`. Dieser Ansatz ermöglicht ein Standard-Hot Backup, bei dem sich die Datenbank im Hot Backup-Modus befindet, gefolgt von einem Snapshot von `zpool1`. Die Datenbank wird dann aus dem Hot Backup-Modus entfernt, das Protokollarchiv wird erzwungen und ein Snapshot von `zpool2` Wird erstellt. Ein Wiederherstellungsvorgang erfordert das Abhängen der zfs-Dateisysteme und den vollständigen Offlining des zpool nach einer SnapRestore-Wiederherstellung. Der zpool kann dann wieder online gebracht werden und die Datenbank wiederhergestellt werden.

Filesystemio_options

Der Oracle-Parameter `filesystemio_options` Funktioniert anders mit ZFS. Wenn `setall` Oder `directio` Wird verwendet, Schreibvorgänge sind synchron und umgehen den BS-Puffer-Cache, aber Lesevorgänge werden von ZFS gepuffert. Diese Aktion führt zu Schwierigkeiten bei der Performance-Analyse, da I/O manchmal vom ZFS-Cache abgefangen und gewartet wird. Dadurch werden die Speicherlatenz und der gesamte I/O geringer als möglicherweise angezeigt.

Netzwerkkonfiguration auf AFF/ FAS -Systemen

Logische Schnittstellen

Oracle Datenbanken benötigen Zugriff auf den Storage. Logische Schnittstellen (LIFs) sind die Netzwerk-Rohrleitungen, die eine Storage Virtual Machine (SVM) mit dem Netzwerk und damit der Datenbank verbinden. Ein angemessenes LIF-Design ist erforderlich, um sicherzustellen, dass für jeden Datenbank-Workload ausreichend Bandbreite vorhanden ist. Das Failover führt nicht zu einem Verlust von Storage-Services.

Dieser Abschnitt bietet einen Überblick über die wichtigsten LIF-Designprinzipien. Eine ausführlichere Dokumentation finden Sie im "Dokumentation zum ONTAP-Netzwerkmanagement". Wie andere Aspekte der Datenbankarchitektur hängen die besten Optionen für die Storage Virtual Machine (SVM, in der CLI als vServer bezeichnet) und das Design der logischen Schnittstelle (LIF) stark von den Skalierungsanforderungen und den geschäftlichen Anforderungen ab.

Berücksichtigen Sie bei der Entwicklung einer LIF-Strategie die folgenden primären Themen:

- **Leistung.** ist die Netzwerkbandbreite ausreichend?
- **Ausfallsicherheit.** gibt es Single Points of Failure im Design?
- **Verwaltbarkeit.** kann das Netzwerk unterbrechungsfrei skaliert werden?

Diese Themen beziehen sich auf die End-to-End-Lösung, vom Host über die Switches bis zum Speichersystem.

LIF-Typen

Es gibt mehrere LIF-Typen. "ONTAP-Dokumentation zu LIF-Typen" Stellen Sie umfassendere Informationen zu diesem Thema bereit, LIFs können jedoch aus funktionaler Sicht in die folgenden Gruppen unterteilt werden:

- **Cluster- und Node-Management-LIFs.** LIFs, die zum Verwalten des Storage-Clusters verwendet werden.
- **SVM-Management-LIFs.** Schnittstellen, die den Zugriff auf eine SVM über die REST-API oder ONTAPI (auch bekannt als ZAPI) für Funktionen wie Snapshot-Erstellung oder Volume-Anpassung erlauben. Produkte wie SnapManager für Oracle (SMO) müssen Zugriff auf eine SVM-Management-LIF haben.
- **Daten-LIFs.** Schnittstellen für FC, iSCSI, NVMe/FC, NVMe/TCP, NFS, oder SMB/CIFS-Daten.

> Eine logische Datenschnittstelle für NFS-Datenverkehr kann durch Änderung der Firewallrichtlinie von auch zum Management verwendet werden `data` Bis `mgmt` Oder eine andere Richtlinie, die HTTP, HTTPS oder SSH erlaubt. Diese Änderung kann die Netzwerkkonfiguration vereinfachen, indem die Konfiguration jedes Hosts für den Zugriff auf die LIF der NFS-Daten und eine separate Management-LIF vermieden wird. Es ist nicht möglich, eine Schnittstelle für iSCSI- und Managementverkehr zu konfigurieren, obwohl beide ein IP-Protokoll verwenden. In iSCSI-Umgebungen ist eine separate Management-LIF erforderlich.

Design von SAN LIF

Das LIF-Design in einer SAN-Umgebung ist aus einem Grund relativ einfach: Multipathing. Alle modernen SAN-Implementierungen ermöglichen es einem Client, über mehrere unabhängige Netzwerkpfade auf Daten zuzugreifen und den optimalen Pfad oder die besten Pfade für den Zugriff auszuwählen. So lässt sich die Performance in Bezug auf LIF-Design einfacher bewältigen, da SAN-Clients automatisch den I/O-Lastausgleich über die besten verfügbaren Pfade durchführen.

Wenn ein Pfad nicht mehr verfügbar ist, wählt der Client automatisch einen anderen Pfad aus. Das daraus resultierende einfache Design macht SAN LIFs im Allgemeinen einfacher zu managen. Das bedeutet nicht, dass eine SAN-Umgebung immer einfacher zu managen ist, da viele andere Aspekte des SAN-Storage viel komplizierter sind als NFS. Es bedeutet schlichtweg, dass das LIF-Design von SAN einfacher ist.

Leistung

Der wichtigste Aspekt bei der LIF-Performance in einer SAN-Umgebung ist die Bandbreite. In einem ONTAP AFF-Cluster mit zwei Nodes mit zwei 16-GB-FC-Ports pro Node können beispielsweise bis zu 32 GB Bandbreite von jedem Node bereitgestellt werden.

Ausfallsicherheit

SAN LIFs führen keinen Failover auf einem AFF Storage-System durch. Wenn eine SAN-LIF aufgrund eines Controller-Failovers ausfällt, erkennt die Multipathing-Software des Clients den Verlust eines Pfads und leitet den I/O an eine andere LIF um. Bei ASA Storage-Systemen wird für LIFs nach kurzer Verzögerung ein Failover durchgeführt. Die I/O wird jedoch nicht unterbrochen, da auf dem anderen Controller bereits aktive Pfade vorhanden sind. Der Failover-Prozess erfolgt, um den Hostzugriff auf allen definierten Ports wiederherzustellen.

Managebarkeit

Die LIF-Migration ist in einer NFS-Umgebung viel üblicher, da die LIF-Migration häufig mit dem Verschieben

von Volumes innerhalb des Clusters verknüpft ist. Wenn Volumes innerhalb des HA-Paars verschoben werden, ist keine Migration einer LIF in eine SAN-Umgebung erforderlich. Der Grund dafür ist, dass ONTAP nach Abschluss der Volume-Verschiebung eine Benachrichtigung über eine Pfadänderung an das SAN sendet, die die SAN-Clients automatisch neu optimieren. Die LIF-Migration mit SAN steht in erster Linie in Verbindung mit größeren Änderungen an physischer Hardware. Wenn beispielsweise ein unterbrechungsfreies Upgrade der Controller erforderlich ist, wird eine SAN LIF auf die neue Hardware migriert. Wenn ein FC-Port defekt ist, kann eine LIF zu einem nicht verwendeten Port migriert werden.

Designempfehlungen

NetApp gibt die folgenden Empfehlungen:

- Erstellen Sie nicht mehr Pfade, als erforderlich sind. Eine übermäßige Anzahl von Pfaden erschwert das gesamte Management und kann zu Problemen mit dem Pfad-Failover auf einigen Hosts führen. Darüber hinaus weisen einige Hosts unerwartete Pfadeinschränkungen für Konfigurationen wie das Booten von SAN auf.
- Nur sehr wenige Konfigurationen sollten mehr als vier Pfade zu einem LUN erfordern. Der Wert von mehr als zwei Nodes, um LUNs bekannt zu machen, ist beschränkt, da das Aggregat, das eine LUN hostet, nicht zugänglich ist, wenn der Node, der Eigentümer der LUN und dessen HA-Partner ausfällt. In solch einem Szenario ist es nicht hilfreich, Pfade auf anderen Nodes als dem primären HA-Paar zu erstellen.
- Obwohl die Anzahl der sichtbaren LUN-Pfade durch Auswählen der in FC-Zonen enthaltenen Ports gemanagt werden kann, ist es im Allgemeinen einfacher, alle potenziellen Zielpunkte in die FC-Zone aufzunehmen und die LUN-Sichtbarkeit auf ONTAP-Ebene zu kontrollieren.
- In ONTAP 8.3 und höher ist die Funktion für die selektive LUN-Zuordnung (SLM) die Standardeinstellung. Bei SLM wird jede neue LUN automatisch von dem Node bereitgestellt, dem das zugrunde liegende Aggregat und der HA-Partner des Node gehören. Durch diese Anordnung müssen keine Portsätze erstellt oder Zoning konfiguriert werden, um den Zugriff auf den Port zu beschränken. Jede LUN ist mit der Mindestanzahl der Nodes verfügbar, die für eine optimale Performance und Stabilität erforderlich sind. *Falls eine LUN außerhalb der beiden Controller migriert werden muss, können die zusätzlichen Knoten mit dem hinzugefügt werden `lun mapping add-reporting-nodes` Befehl, sodass die LUNs auf den neuen Nodes angekündigt werden. Dadurch werden zusätzliche SAN-Pfade zu den LUNs für die LUN-Migration erstellt. Der Host muss jedoch einen Erkennungsvorgang durchführen, um die neuen Pfade verwenden zu können.
- Seien Sie nicht übermäßig besorgt über indirekten Verkehr. Es empfiehlt sich, in einer sehr I/O-intensiven Umgebung, in der jede Mikrosekunde von großer Latenz ist, indirekten Verkehr zu vermeiden, aber der sichtbare Performance-Effekt ist bei typischen Workloads zu vernachlässigen.

LIF-Design von NFS

Im Gegensatz zu SAN-Protokollen kann bei NFS nur bedingt mehrere Pfade zu Daten definiert werden. Die parallelen NFS-Erweiterungen (pNFS) zu NFSv4 beheben diese Einschränkung. Da die ethernet-Geschwindigkeit jedoch 100 GB erreicht hat, ist das Hinzufügen weiterer Pfade selten ein Nutzen.

Performance und Ausfallsicherheit

Obwohl die Messung der LIF-Performance in erster Linie dazu dient, die gesamte Bandbreite von allen primären Pfaden zu berechnen, muss die Bestimmung der Performance von NFS LIF genau die Netzwerkkonfiguration durchgeführt werden. Beispielsweise können zwei 10-Gbit-Ports als physische Rohports konfiguriert oder als LACP-Interface-Gruppe (Link Aggregation Control Protocol) konfiguriert werden. Wenn sie als Schnittstellengruppe konfiguriert sind, stehen mehrere Load-Balancing-Richtlinien zur Verfügung, die je nachdem, ob der Datenverkehr geswitcht oder geroutet wird, unterschiedlich funktionieren. Oracle Direct NFS (dNFS) bietet Load-Balancing-Konfigurationen, die derzeit in keinen NFS-Clients des Betriebssystems vorhanden sind.

Im Gegensatz zu SAN-Protokollen erfordern NFS-Filesysteme Ausfallsicherheit auf Protokollebene. Beispielsweise wird eine LUN immer mit aktiviertem Multipathing konfiguriert, was bedeutet, dass dem Storage-System mehrere redundante Kanäle zur Verfügung stehen, von denen jeder das FC-Protokoll verwendet. Ein NFS-Dateisystem hingegen hängt von der Verfügbarkeit eines einzelnen TCP/IP-Kanals ab, der nur auf der physischen Ebene geschützt werden kann. Diese Anordnung ist, warum Optionen wie Port-Failover und LACP Port-Aggregation existieren.

In einer NFS-Umgebung werden sowohl Performance als auch Ausfallsicherheit auf der Netzwerkprotokollebene bereitgestellt. Dadurch sind beide Themen miteinander verflochten und müssen gemeinsam diskutiert werden.

Binden Sie LIFs an Portgruppen

Um ein LIF an eine Portgruppe zu binden, ordnen Sie die LIF-IP-Adresse einer Gruppe physischer Ports zu. Die primäre Methode zur Aggregation physischer Ports ist LACP. Die Fehlertoleranz-Funktion von LACP ist ziemlich einfach. Jeder Port in einer LACP-Gruppe wird überwacht und im Falle einer Störung aus der Portgruppe entfernt. Es gibt jedoch viele Missverständnisse darüber, wie LACP in Bezug auf Performance funktioniert:

- Für LACP ist keine Konfiguration auf dem Switch erforderlich, um mit dem Endpunkt übereinstimmen zu können. Beispielsweise kann ONTAP mit IP-basiertem Lastausgleich konfiguriert werden, während ein Switch MAC-basierten Lastausgleich verwenden kann.
- Jeder Endpunkt, der eine LACP-Verbindung verwendet, kann den Port für die Paketübertragung unabhängig auswählen, jedoch nicht den für den Empfang verwendeten Port auswählen. Das bedeutet, dass Datenverkehr von ONTAP zu einem bestimmten Ziel an einen bestimmten Port gebunden ist, und der Rückverkehr könnte auf einer anderen Schnittstelle eintreffen. Dies verursacht jedoch keine Probleme.
- LACP verteilt den Datenverkehr nicht ständig gleichmäßig. In einer großen Umgebung mit vielen NFS-Clients wird normalerweise sogar alle Ports in einer LACP-Aggregation genutzt. Jedoch ist jedes ein NFS-Dateisystem in der Umgebung auf die Bandbreite von nur einem Port beschränkt, nicht die gesamte Aggregation.
- Obwohl LACP-Richtlinien für die Robin-Lösung auf ONTAP verfügbar sind, adressieren diese Richtlinien nicht die Verbindung von einem Switch zu einem Host. Beispielsweise ist eine Konfiguration mit einem LACP Trunk mit vier Ports auf einem Host und einem LACP Trunk mit vier Ports auf einem ONTAP immer noch nur in der Lage, ein Filesystem über einen einzelnen Port zu lesen. Obwohl ONTAP Daten über alle vier Ports übertragen kann, sind derzeit keine Switch-Technologien verfügbar, die über alle vier Ports vom Switch an den Host gesendet werden. Es wird nur eine verwendet.

In größeren Umgebungen, die aus vielen Datenbank-Hosts bestehen, ist der geläufigste Ansatz, mithilfe eines IP-Lastausgleichs ein LACP Aggregat mit einer entsprechenden Anzahl von 10 GB (oder schneller) Schnittstellen zu erstellen. Mit diesem Ansatz kann ONTAP sogar die Nutzung aller Ports ermöglichen, sofern genügend Clients vorhanden sind. Der Lastausgleich wird unterbrochen, wenn weniger Clients in der Konfiguration vorhanden sind, da LACP Trunking die Last nicht dynamisch neu verteilt.

Wenn eine Verbindung hergestellt wird, wird der Datenverkehr in eine bestimmte Richtung nur an einem Port platziert. Beispielsweise liest eine Datenbank, die einen vollständigen Tabellenscan gegen ein NFS-Dateisystem durchführt, das über einen LACP-Trunk mit vier Ports verbunden ist, Daten über nur eine Netzwerkkarte (NIC). Wenn sich nur drei Datenbankserver in einer solchen Umgebung befinden, ist es möglich, dass alle drei vom gleichen Port lesen, während die anderen drei Ports inaktiv sind.

Binden Sie LIFs an physische Ports

Das Binden einer LIF an einen physischen Port führt zu einer granulareren Kontrolle der Netzwerkkonfiguration, da eine gegebene IP-Adresse auf einem ONTAP-System jeweils nur mit einem

Netzwerk-Port verknüpft ist. Stabilität wird dann durch die Konfiguration von Failover-Gruppen und Failover-Richtlinien erreicht.

Failover-Richtlinien und Failover-Gruppen

Das Verhalten von LIFs wird während der Netzwerkunterbrechung durch Failover-Richtlinien und Failover-Gruppen gesteuert. Die Konfigurationsoptionen wurden mit den verschiedenen Versionen von ONTAP geändert. Konsultieren Sie die "ONTAP Netzwerkmanagement-Dokumentation für Failover-Gruppen und Richtlinien" Finden Sie spezifische Details zur implementierten Version von ONTAP.

ONTAP 8.3 und höher ermöglichen das Management von LIF-Failovers basierend auf Broadcast-Domänen. Daher kann ein Administrator alle Ports definieren, die Zugriff auf ein bestimmtes Subnetz haben, und ONTAP erlauben, eine entsprechende Failover-LIF auszuwählen. Einige Kunden verwenden diesen Ansatz durchaus, weist jedoch aufgrund der mangelnden Planbarkeit in einer Storage-Netzwerkumgebung mit hoher Geschwindigkeit Einschränkungen auf. Beispielsweise kann eine Umgebung sowohl 1-Gbit-Ports für routinemäßigen Filesystem-Zugriff als auch 10-Gbit-Ports für Datendatei-I/O. Wenn beide Ports in derselben Broadcast-Domäne vorhanden sind, kann ein LIF-Failover dazu führen, Datendatei-I/O von einem 10-GB-Port auf einen 1-GB-Port zu verschieben.

Zusammenfassend lassen sich die folgenden Vorgehensweisen berücksichtigen:

1. Konfigurieren Sie eine Failover-Gruppe als benutzerdefiniert.
2. Füllen Sie die Failover-Gruppe mit Ports am Partner-Controller für Storage Failover (SFO), damit die LIFs beim Storage Failover den Aggregaten folgen. Dadurch wird die Erstellung indirekter Verkehrsströme vermieden.
3. Verwenden Sie Failover-Ports, deren Performance-Merkmale mit der ursprünglichen logischen Schnittstelle übereinstimmen. Beispielsweise sollte eine LIF auf einem einzelnen physischen 10-Gbit-Port eine Failover-Gruppe mit einem einzelnen 10-Gbit-Port enthalten. Ein LACP LIF mit vier Ports sollte ein Failover auf eine andere LACP LIF mit vier Ports durchführen. Diese Ports wären eine Teilmenge der Ports, die in der Broadcast-Domäne definiert sind.
4. Setzen Sie die Failover-Richtlinie auf nur SFO-Partner. Dadurch wird sichergestellt, dass die LIF während des Failovers dem Aggregat folgt.

Autom. Rücksetzung

Stellen Sie die ein `auto-revert` Parameter wie gewünscht. Die meisten Kunden bevorzugen es, diesen Parameter auf zu setzen `true` Um das LIF auf seinen Home Port zurückzusetzen. In einigen Fällen haben Kunden dies jedoch auf `false `so gesetzt, dass ein unerwartetes Failover untersucht werden kann, bevor eine LIF an ihren Home Port zurückgegeben wird.

LIF-Volume-Verhältnis

Ein weit verbreitetes Missverständnis ist, dass es eine 1:1 Beziehung zwischen Volumes und NFS LIFs geben muss. Diese Konfiguration ist zwar erforderlich, um ein Volume ohne zusätzlichen Interconnect-Verkehr an eine beliebige Stelle in einem Cluster zu verschieben, ist jedoch kategorisch keine Anforderung. Der Intercluster-Datenverkehr muss berücksichtigt werden, aber die bloße Anwesenheit von Intercluster-Datenverkehr verursacht keine Probleme. Viele der für ONTAP veröffentlichten Benchmarks sind überwiegend indirekte I/O-Vorgänge

Ein Datenbankprojekt mit einer relativ kleinen Anzahl Performance-kritischer Datenbanken, für die nur insgesamt 40 Volumes benötigt wurden, könnte beispielsweise eine LIF-Strategie für das 1:1 Volume rechtfertigen. Dieses Arrangement würde 40 IP-Adressen erfordern. Jedes Volume könnte dann zusammen mit der zugehörigen LIF an jeden beliebigen Ort im Cluster verschoben werden. Der Datenverkehr würde dann

immer direkt erfolgen, wodurch jede Latenzquelle sogar auf Mikrosekunden-Ebene minimiert wird.

Zählerbeispiel: Eine große, gehostete Umgebung kann durch eine 1:1:1-Beziehung zwischen Kunden und LIFs einfacher gemanagt werden. Im Laufe der Zeit muss ein Volume möglicherweise auf einen anderen Node migriert werden, was zu einem indirekten Traffic führen würde. Der Performance-Effekt sollte jedoch nicht nachweisbar sein, es sei denn, die Netzwerk-Ports auf dem Interconnect-Switch sind voll ausgelastet. Falls Bedenken bestehen, kann eine neue LIF auf zusätzlichen Nodes erstellt werden, und der Host kann im nächsten Wartungsfenster aktualisiert werden, um indirekten Traffic aus der Konfiguration zu entfernen.

TCP/IP- und ethernet-Konfiguration

Viele Kunden von Oracle auf ONTAP verwenden ethernet, das Netzwerkprotokoll von NFS, iSCSI, NVMe/TCP sowie insbesondere die Cloud.

Einstellungen für das Host-Betriebssystem

Die Dokumentation der meisten Anwendungsanbieter enthält bestimmte TCP- und ethernet-Einstellungen, die sicherstellen sollen, dass die Anwendung optimal funktioniert. Diese Einstellungen reichen in der Regel aus, um auch eine optimale IP-basierte Speicherleistung zu erzielen.

Ethernet-Flusskontrolle

Mit dieser Technologie kann ein Client verlangen, dass ein Sender die Datenübertragung vorübergehend stoppt. Dies geschieht normalerweise, weil der Empfänger eingehende Daten nicht schnell genug verarbeiten kann. Die Anforderung, dass ein Sender die Übertragung abbricht, war zu einem Zeitpunkt weniger störend, als dass ein Empfänger Pakete verwirft, weil die Puffer voll waren. Dies ist bei den heute in Betriebssystemen verwendeten TCP-Stacks nicht mehr der Fall. Tatsächlich verursacht die Flusskontrolle mehr Probleme als sie löst.

Leistungsprobleme, die durch die Ethernet-Flusssteuerung verursacht werden, haben in den letzten Jahren zugenommen. Der Grund dafür ist, dass die Ethernet-Flusssteuerung auf der physischen Ebene ausgeführt wird. Wenn eine Netzwerkkonfiguration es einem Host-Betriebssystem ermöglicht, eine Ethernet-Datenflusssteuerungsanforderung an ein Storage-System zu senden, führt dies zu einer I/O-Pause für alle verbundenen Clients. Da immer mehr Clients von einem einzelnen Storage Controller bedient werden, steigt die Wahrscheinlichkeit, dass ein oder mehrere dieser Clients Flow Control-Anfragen senden. Das Problem ist bei Kundenstandorten mit umfassender Betriebssystemvirtualisierung häufig aufgetreten.

Eine NIC auf einem NetApp-System sollte keine Anfragen zur Flusskontrolle empfangen. Die Methode, mit der dieses Ergebnis erzielt wird, hängt vom Hersteller des Netzwerk-Switches ab. In den meisten Fällen kann die Flusssteuerung auf einem Ethernet-Switch eingestellt werden `receive desired` Oder `receive on`, Das bedeutet, dass eine Durchflussregelanforderung nicht an den Speichercontroller weitergeleitet wird. In anderen Fällen lässt die Netzwerkverbindung auf dem Storage Controller möglicherweise die Deaktivierung der Flusssteuerung nicht zu. In diesen Fällen müssen die Clients so konfiguriert werden, dass sie keine Flow-Control-Anforderungen senden, entweder indem sie auf die NIC-Konfiguration auf dem Host-Server selbst oder auf die Switch-Ports wechseln, mit denen der Host-Server verbunden ist.

NetApp empfiehlt sicherzustellen, dass NetApp-Speicher-Controller keine Ethernet-Flow-Control-Pakete empfangen. Dies kann im Allgemeinen durch Einstellen der Switch Ports geschehen, an die der Controller angeschlossen ist. Bei einigen Switch-Hardware bestehen jedoch Einschränkungen, die stattdessen clientseitige Änderungen erfordern können.

MTU-Größen

Der Einsatz von Jumbo Frames hat gezeigt, dass sich die Performance in 1-GB-Netzwerken durch Reduzierung des CPU- und Netzwerk-Overheads verbessert. Die Vorteile sind jedoch in der Regel nicht signifikant.

NetApp empfiehlt, wenn möglich Jumbo Frames zu implementieren, sowohl um potenzielle Leistungsvorteile zu realisieren als auch um die Lösung zukunftssicher zu machen.

Die Verwendung von Jumbo Frames in einem 10-Gbit-Netzwerk ist fast zwingend erforderlich. Der Grund dafür ist, dass die meisten 10-GB-Implementierungen vor Erreichen der 10-GB-Marke ohne Jumbo-Frames eine Grenze von Paketen pro Sekunde erreichen. Die Verwendung von Jumbo Frames verbessert die Effizienz bei der TCP/IP-Verarbeitung, da Betriebssystem, Server, NICs und Speichersystem weniger, aber größere Pakete verarbeiten können. Die Leistungsverbesserung variiert von NIC zu NIC, ist jedoch signifikant.

Bei Jumbo-Frame-Implementierungen besteht die allgemeine, aber falsche Annahme, dass alle verbundenen Geräte Jumbo-Frames unterstützen müssen und dass die MTU-Größe End-to-End entsprechen muss Stattdessen verhandeln die beiden Netzwerkendpunkte beim Herstellen einer Verbindung die höchste für beide Seiten akzeptable Frame-Größe. In einer typischen Umgebung ist ein Netzwerk-Switch auf eine MTU-Größe von 9216, der NetApp-Controller auf 9000 und die Clients auf 9000 und 1514 eingestellt. Clients, die eine MTU von 9000 unterstützen, können Jumbo-Frames verwenden, und Clients, die nur 1514 unterstützen, können einen niedrigeren Wert aushandeln.

Probleme mit dieser Anordnung sind in einer komplett geschalteten Umgebung selten. Achten Sie jedoch in einer gerouteten Umgebung darauf, dass kein Zwischenrouter gezwungen ist, Jumbo-Frames zu fragmentieren.

NetApp empfiehlt die Konfiguration folgender Komponenten:

- Jumbo Frames sind wünschenswert, jedoch nicht erforderlich mit 1Gb Ethernet (GbE).
- Jumbo Frames sind für maximale Performance mit 10 GbE und schneller erforderlich.

TCP-Parameter

Drei Einstellungen sind oft falsch konfiguriert: TCP-Zeitstempel, selektive Bestätigung (SACK) und TCP-Fenster-Skalierung. Viele veraltete Dokumente im Internet empfehlen, einen oder mehrere dieser Parameter zu deaktivieren, um die Leistung zu verbessern. Vor vielen Jahren war diese Empfehlung verdienlich, als die CPU-Kapazitäten wesentlich geringer waren und der Overhead für die TCP-Verarbeitung, wenn möglich, reduziert werden konnte.

Bei modernen Betriebssystemen führt die Deaktivierung dieser TCP-Funktionen jedoch in der Regel nicht zu nachweisbaren Vorteilen und kann gleichzeitig die Leistung beeinträchtigen. In virtualisierten Netzwerkumgebungen sind Performance-Schäden besonders wahrscheinlich, da diese Funktionen für eine effiziente Handhabung von Paketverlusten und Änderungen der Netzwerkqualität erforderlich sind.

NetApp empfiehlt, TCP-Zeitstempel, SACK und TCP-Fenster-Skalierung auf dem Host zu aktivieren, und alle drei dieser Parameter sollten in jedem aktuellen Betriebssystem standardmäßig aktiviert sein.

FC SAN-Konfiguration

Bei der Konfiguration von FC SAN für Oracle-Datenbanken geht es in erster Linie um die

Umsetzung der täglichen SAN Best Practices.

Dazu gehören typische Planungsmaßnahmen wie die Sicherstellung einer ausreichenden Bandbreite auf dem SAN zwischen dem Host und dem Speichersystem, die Überprüfung, ob alle SAN-Pfade zwischen allen erforderlichen Geräten vorhanden sind, unter Verwendung der FC-Port-Einstellungen, die Ihr FC-Switch-Anbieter benötigt, um ISL-Konflikte zu vermeiden, und ordnungsgemäße Überwachung des SAN-Fabrics verwenden.

Zoning

Eine FC-Zone sollte nie mehr als einen Initiator enthalten. Eine solche Anordnung mag zunächst zu funktionieren scheinen, doch Crosstalk zwischen Initiatoren beeinträchtigt letztendlich die Performance und Stabilität.

Multitarget-Zonen werden allgemein als sicher angesehen, obwohl in seltenen Fällen das Verhalten von FC-Zielports unterschiedlicher Anbieter Probleme verursacht hat. Es ist beispielsweise zu vermeiden, die Ziel-Ports von einem NetApp und einem nicht-NetApp Storage-Array in derselben Zone zu integrieren. Darüber hinaus besteht mit noch größerer Wahrscheinlichkeit die Gefahr, dass ein NetApp Storage-System und ein Bandgerät in dieselbe Zone platziert werden.

Direct-Connect-Netzwerk

Storage-Administratoren ziehen es manchmal vor, ihre Infrastruktur zu vereinfachen, indem sie Netzwerk-Switches von der Konfiguration entfernen. Dies kann in einigen Szenarien unterstützt werden.

ISCSI und NVMe/TCP

Ein Host, der iSCSI oder NVMe/TCP verwendet, kann direkt mit einem Storage-System verbunden werden und ordnungsgemäß ausgeführt werden. Der Grund dafür ist Pathing. Direkte Verbindungen zu zwei verschiedenen Storage Controllern ergeben zwei unabhängige Pfade für den Datenfluss. Der Verlust von Pfad, Port oder Controller verhindert nicht, dass der andere Pfad verwendet wird.

NFS

Direct-Connected NFS Storage kann genutzt werden, aber mit einer erheblichen Einschränkung - Failover funktioniert nicht ohne einen erheblichen Scripting-Aufwand, der in der Verantwortung des Kunden liegt.

Der Grund, warum ein unterbrechungsfreier Failover mit direkt verbundenem NFS-Storage kompliziert ist, ist das Routing auf dem lokalen Betriebssystem. Angenommen, ein Host hat eine IP-Adresse von 192.168.1.1/24 und ist direkt mit einem ONTAP-Controller mit einer IP-Adresse von 192.168.1.50/24 verbunden. Während eines Failovers kann diese 192.168.1.50-Adresse ein Failover auf den anderen Controller durchführen, und sie wird für den Host verfügbar sein. Wie erkennt der Host jedoch sein Vorhandensein? Die ursprüngliche 192.168.1.1-Adresse ist noch auf der Host-NIC vorhanden, die keine Verbindung mehr zu einem Betriebssystem herstellt. Der für 192.168.1.50 bestimmte Datenverkehr würde weiterhin an einen nicht funktionsfähigen Netzwerkport gesendet.

Die zweite BS-NIC könnte als 19 konfiguriert werden 2.168.1.2 und wäre in der Lage, mit der Failed Over 192.168.1.50-Adresse zu kommunizieren, aber die lokalen Routing-Tabellen würden standardmäßig eine **und nur eine**-Adresse verwenden, um mit dem Subnetz 192.168.1.0/24 zu kommunizieren. Ein Sysadmin könnte ein Skript-Framework erstellen, das eine fehlerhafte Netzwerkverbindung erkennt und die lokalen Routing-Tabellen ändert oder Schnittstellen hoch- und herunterfahren würde. Das genaue Verfahren hängt vom verwendeten Betriebssystem ab.

In der Praxis haben NetApp-Kunden NFS direkt verbunden, aber normalerweise nur für Workloads, bei denen IO-Pausen während Failover akzeptabel sind. Wenn harte Mounts verwendet werden, sollte es während solcher Pausen keine IO-Fehler geben. Die E/A-Vorgänge sollten so lange hängen bleiben, bis Dienste wiederhergestellt werden, entweder durch ein Failback oder durch einen manuellen Eingriff, um IP-Adressen zwischen NICs auf dem Host zu verschieben.

FC-Direktverbindung

Es ist nicht möglich, einen Host direkt über das FC-Protokoll mit einem ONTAP Storage-System zu verbinden. Der Grund dafür ist die Verwendung von NPIV. Der WWN, der einen ONTAP FC-Port mit dem FC-Netzwerk identifiziert, verwendet eine Art Virtualisierung, die als NPIV bezeichnet wird. Jedes Gerät, das an ein ONTAP-System angeschlossen ist, muss einen NPIV-WWN erkennen können. Es gibt derzeit keine HBA-Anbieter, die einen HBA anbieten, der auf einem Host installiert werden kann, der ein NPIV-Ziel unterstützen könnte.

Netzwerkkonfiguration auf ASA r2-Systemen

Logische Schnittstellen

Oracle Datenbanken benötigen Zugriff auf den Storage. Logische Schnittstellen (LIFs) sind die Netzwerk-Rohrleitungen, die eine Storage Virtual Machine (SVM) mit dem Netzwerk und damit der Datenbank verbinden. Ein angemessenes LIF-Design ist erforderlich, um sicherzustellen, dass für jeden Datenbank-Workload ausreichend Bandbreite vorhanden ist. Das Failover führt nicht zu einem Verlust von Storage-Services.

Dieser Abschnitt bietet einen Überblick über die wichtigsten LIF-Designprinzipien für ASA r2-Systeme, die für reine SAN-Umgebungen optimiert sind. Eine ausführlichere Dokumentation finden Sie unter "Dokumentation zum ONTAP-Netzwerkmanagement"Die Wie bei anderen Aspekten der Datenbankarchitektur hängen die besten Optionen für die Gestaltung von Storage Virtual Machines (SVM, in der CLI als vserver bezeichnet) und Logical Interfaces (LIF) stark von den Skalierungsanforderungen und den Geschäftsbedürfnissen ab.

Berücksichtigen Sie bei der Entwicklung einer LIF-Strategie die folgenden primären Themen:

- **Leistung.** Ist die Netzwerkbandbreite für Oracle-Workloads ausreichend?
- **Ausfallsicherheit.** gibt es Single Points of Failure im Design?
- **Verwaltbarkeit.** kann das Netzwerk unterbrechungsfrei skaliert werden?

Diese Themen beziehen sich auf die End-to-End-Lösung, vom Host über die Switches bis zum Speichersystem.

LIF-Typen

Es gibt mehrere LIF-Typen. "ONTAP-Dokumentation zu LIF-Typen" Stellen Sie umfassendere Informationen zu diesem Thema bereit, LIFs können jedoch aus funktionaler Sicht in die folgenden Gruppen unterteilt werden:

- **Cluster- und Node-Management-LIFs.** LIFs, die zum Verwalten des Storage-Clusters verwendet werden.
- **SVM-Management-LIFs.** Schnittstellen, die den Zugriff auf eine SVM über die REST-API oder ONTAPI (auch bekannt als ZAPI) für Funktionen wie Snapshot-Erstellung oder Volume-Anpassung erlauben. Produkte wie SnapManager für Oracle (SMO) müssen Zugriff auf eine SVM-Management-LIF haben.
- **Daten-LIFs.** Schnittstellen nur für SAN-Protokolle: FC, iSCSI, NVMe/FC, NVMe/TCP. NAS-Protokolle (NFS, SMB/CIFS) werden auf ASA r2-Systemen nicht unterstützt.

Es ist nicht möglich, eine Schnittstelle sowohl für iSCSI (oder NVMe/TCP) als auch für Management-Datenverkehr zu konfigurieren, obwohl beide ein IP-Protokoll verwenden. In iSCSI- oder NVMe/TCP-Umgebungen ist ein separates Management-LIF erforderlich. Zur Gewährleistung von Ausfallsicherheit und Leistung konfigurieren Sie mehrere SAN-Daten-LIFs pro Protokoll und Knoten und verteilen Sie diese auf verschiedene physische Ports und Fabrics. Im Gegensatz zu AFF/ FAS -Systemen erlaubt ASA r2 keinen NFS- oder SMB-Datenverkehr, daher gibt es keine Möglichkeit, eine NAS-Daten-LIF für die Verwaltung umzufunktionieren.

Design von SAN LIF

Das LIF-Design in einer SAN-Umgebung ist aus einem Grund relativ einfach: Multipathing. Alle modernen SAN-Implementierungen ermöglichen es einem Client, über mehrere unabhängige Netzwerkpfade auf Daten zuzugreifen und den optimalen Pfad oder die besten Pfade für den Zugriff auszuwählen. So lässt sich die Performance in Bezug auf LIF-Design einfacher bewältigen, da SAN-Clients automatisch den I/O-Lastausgleich über die besten verfügbaren Pfade durchführen.

Wenn ein Pfad nicht mehr verfügbar ist, wählt der Client automatisch einen anderen Pfad aus. Das daraus resultierende einfache Design macht SAN LIFs im Allgemeinen einfacher zu managen. Das bedeutet nicht, dass eine SAN-Umgebung immer einfacher zu managen ist, da viele andere Aspekte des SAN-Storage viel komplizierter sind als NFS. Es bedeutet schlichtweg, dass das LIF-Design von SAN einfacher ist.

Leistung

Der wichtigste Faktor für die Leistungsfähigkeit von LIF in einer SAN-Umgebung ist die Bandbreite. Beispielsweise ermöglicht ein ASA r2-Cluster mit zwei Knoten und zwei 32-Gb-FC-Ports pro Knoten eine Bandbreite von bis zu 64 Gb zu/von jedem Knoten. Stellen Sie für NVMe/TCP oder iSCSI ebenfalls sicher, dass für Oracle-Workloads ausreichend 25GbE- oder 100GbE-Konnektivität vorhanden ist.

Ausfallsicherheit

SAN-LIFs funktionieren nicht auf die gleiche Weise wie NAS-LIFs. ASA r2-Systeme nutzen Host-Multipathing (MPIO/ALUA) für Ausfallsicherheit. Wenn ein SAN LIF aufgrund eines Controller-Failovers nicht verfügbar ist, erkennt die Multipathing-Software des Clients den Verlust eines Pfades und leitet die E/A auf einen alternativen Pfad um. ASA r2 kann nach einer kurzen Verzögerung eine LIF-Verschiebung durchführen, um die volle Pfadverfügbarkeit wiederherzustellen. Dies unterbricht jedoch nicht die E/A, da auf dem Partnerknoten bereits aktive Pfade vorhanden sind. Der Failover-Prozess dient dazu, den Hostzugriff auf alle definierten Ports wiederherzustellen.

Managebarkeit

Eine Migration einer LIF in einer SAN-Umgebung ist nicht erforderlich, wenn Volumes innerhalb des HA-Paares verschoben werden. Das liegt daran, dass ONTAP nach Abschluss der Volume-Verschiebung eine Benachrichtigung an das SAN über eine Pfadänderung sendet und die SAN-Clients diese automatisch neu optimieren. Die Migration von LIF zu SAN ist in erster Linie mit größeren physischen Hardwareänderungen verbunden. Wenn beispielsweise ein Upgrade der Controller ohne Betriebsunterbrechung erforderlich ist, wird ein SAN LIF auf die neue Hardware migriert. Falls sich ein FC-Port als fehlerhaft erweist, kann ein LIF auf einen ungenutzten Port migriert werden.

Designempfehlungen

NetApp gibt für ASA r2 SAN-Umgebungen folgende Empfehlungen:

- Erstellen Sie nicht mehr Pfade, als erforderlich sind. Eine übermäßige Anzahl von Pfaden erschwert das gesamte Management und kann zu Problemen mit dem Pfad-Failover auf einigen Hosts führen. Darüber

hinaus weisen einige Hosts unerwartete Pfadeinschränkungen für Konfigurationen wie das Booten von SAN auf.

- Nur sehr wenige Konfigurationen sollten mehr als vier Pfade zu einem LUN erfordern. Der Wert von mehr als zwei Nodes, um LUNs bekannt zu machen, ist beschränkt, da das Aggregat, das eine LUN hostet, nicht zugänglich ist, wenn der Node, der Eigentümer der LUN und dessen HA-Partner ausfällt. In solch einem Szenario ist es nicht hilfreich, Pfade auf anderen Nodes als dem primären HA-Paar zu erstellen.
- Obwohl die Anzahl der sichtbaren LUN-Pfade durch Auswählen der in FC-Zonen enthaltenen Ports gemanagt werden kann, ist es im Allgemeinen einfacher, alle potenziellen Zielpunkte in die FC-Zone aufzunehmen und die LUN-Sichtbarkeit auf ONTAP-Ebene zu kontrollieren.
- Nutzen Sie die Funktion „Selective LUN Mapping“ (SLM), die standardmäßig aktiviert ist. Mit SLM wird jede neue LUN automatisch von dem Knoten, dem das zugrunde liegende Aggregat gehört, und dem HA-Partner des Knotens angekündigt. Durch diese Anordnung entfällt die Notwendigkeit, Portgruppen zu erstellen oder Zonen zu konfigurieren, um die Portzugänglichkeit einzuschränken. Jede LUN ist auf der minimalen Anzahl von Knoten verfügbar, die für optimale Leistung und Ausfallsicherheit erforderlich sind.
- Falls eine LUN außerhalb der beiden Controller migriert werden muss, können die zusätzlichen Knoten mit der `lun mapping add-reporting-nodes` Befehl, damit die LUNs auf den neuen Knoten bekanntgegeben werden. Dadurch werden zusätzliche SAN-Pfade zu den LUNs für die LUN-Migration erstellt. Allerdings muss der Host eine Erkennungsoperation durchführen, um die neuen Pfade nutzen zu können.
- Seien Sie nicht übermäßig besorgt über indirekten Verkehr. Es empfiehlt sich, in einer sehr I/O-intensiven Umgebung, in der jede Mikrosekunde von großer Latenz ist, indirekten Verkehr zu vermeiden, aber der sichtbare Performance-Effekt ist bei typischen Workloads zu vernachlässigen.

TCP/IP- und ethernet-Konfiguration

Viele Oracle on ASA r2 ONTAP Kunden verwenden Ethernet, das Netzwerkprotokoll von iSCSI und NVMe/TCP.

Einstellungen für das Host-Betriebssystem

Die Dokumentation der meisten Anwendungsanbieter enthält bestimmte TCP- und ethernet-Einstellungen, die sicherstellen sollen, dass die Anwendung optimal funktioniert. Diese Einstellungen reichen in der Regel aus, um auch eine optimale IP-basierte Speicherleistung zu erzielen.

Ethernet-Flusskontrolle

Mit dieser Technologie kann ein Client verlangen, dass ein Sender die Datenübertragung vorübergehend stoppt. Dies geschieht normalerweise, weil der Empfänger eingehende Daten nicht schnell genug verarbeiten kann. Die Anforderung, dass ein Sender die Übertragung abbricht, war zu einem Zeitpunkt weniger störend, als dass ein Empfänger Pakete verwirft, weil die Puffer voll waren. Dies ist bei den heute in Betriebssystemen verwendeten TCP-Stacks nicht mehr der Fall. Tatsächlich verursacht die Flusskontrolle mehr Probleme als sie löst.

Leistungsprobleme, die durch die Ethernet-Flusssteuerung verursacht werden, haben in den letzten Jahren zugenommen. Der Grund dafür ist, dass die Ethernet-Flusssteuerung auf der physischen Ebene ausgeführt wird. Wenn eine Netzwerkkonfiguration es einem Host-Betriebssystem ermöglicht, eine Ethernet-Datenflusssteuerungsanforderung an ein Storage-System zu senden, führt dies zu einer I/O-Pause für alle verbundenen Clients. Da immer mehr Clients von einem einzelnen Storage Controller bedient werden, steigt die Wahrscheinlichkeit, dass ein oder mehrere dieser Clients Flow Control-Anfragen senden. Das Problem ist bei Kundenstandorten mit umfassender Betriebssystemvirtualisierung häufig aufgetreten.

Eine NIC auf einem NetApp-System sollte keine Anfragen zur Flusskontrolle empfangen. Die Methode, mit der

dieses Ergebnis erzielt wird, hängt vom Hersteller des Netzwerk-Switches ab. In den meisten Fällen kann die Flusssteuerung auf einem Ethernet-Switch eingestellt werden `receive desired` Oder `receive on`, Das bedeutet, dass eine Durchflussregelanforderung nicht an den Speichercontroller weitergeleitet wird. In anderen Fällen lässt die Netzwerkverbindung auf dem Storage Controller möglicherweise die Deaktivierung der Flusssteuerung nicht zu. In diesen Fällen müssen die Clients so konfiguriert werden, dass sie keine Flow-Control-Anforderungen senden, entweder indem sie auf die NIC-Konfiguration auf dem Host-Server selbst oder auf die Switch-Ports wechseln, mit denen der Host-Server verbunden ist.

Bei ASA r2-Systemen, die ausschließlich für SAN ausgelegt sind, gelten die Überlegungen zur Ethernet-Flusskontrolle primär für iSCSI- und NVMe/TCP-Datenverkehr.

* NetApp empfiehlt* sicherzustellen, dass NetApp ASA r2 Storage-Controller keine Ethernet-Flow-Control-Pakete empfangen. Dies kann im Allgemeinen durch die Konfiguration der Switch-Ports, an die der Controller angeschlossen ist, erfolgen. Bei einigen Switch-Hardware gibt es jedoch Einschränkungen, die möglicherweise clientseitige Änderungen erfordern.

MTU-Größen

Der Einsatz von Jumbo Frames hat gezeigt, dass sich die Performance in 1-GB-Netzwerken durch Reduzierung des CPU- und Netzwerk-Overheads verbessert. Die Vorteile sind jedoch in der Regel nicht signifikant.

NetApp empfiehlt, wenn möglich Jumbo Frames zu implementieren, sowohl um potenzielle Leistungsvorteile zu realisieren als auch um die Lösung zukunftssicher zu machen.

Bei ASA r2-Systemen, die ausschließlich für SAN ausgelegt sind, gelten Jumbo-Frames nur für Ethernet-basierte SAN-Protokolle (iSCSI und NVMe/TCP).

Die Verwendung von Jumbo Frames in einem 10-Gbit-Netzwerk ist fast zwingend erforderlich. Der Grund dafür ist, dass die meisten 10-GB-Implementierungen vor Erreichen der 10-GB-Marke ohne Jumbo-Frames eine Grenze von Paketen pro Sekunde erreichen. Die Verwendung von Jumbo Frames verbessert die Effizienz bei der TCP/IP-Verarbeitung, da Betriebssystem, Server, NICs und Speichersystem weniger, aber größere Pakete verarbeiten können. Die Leistungsverbesserung variiert von NIC zu NIC, ist jedoch signifikant.

Bei Jumbo-Frame-Implementierungen besteht die allgemeine, aber falsche Annahme, dass alle verbundenen Geräte Jumbo-Frames unterstützen müssen und dass die MTU-Größe End-to-End entsprechen muss Stattdessen verhandeln die beiden Netzwerkendpunkte beim Herstellen einer Verbindung die höchste für beide Seiten akzeptable Frame-Größe. In einer typischen Umgebung ist ein Netzwerk-Switch auf eine MTU-Größe von 9216, der NetApp-Controller auf 9000 und die Clients auf 9000 und 1514 eingestellt. Clients, die eine MTU von 9000 unterstützen, können Jumbo-Frames verwenden, und Clients, die nur 1514 unterstützen, können einen niedrigeren Wert aushandeln.

Probleme mit dieser Anordnung sind in einer komplett geschalteten Umgebung selten. Achten Sie jedoch in einer gerouteten Umgebung darauf, dass kein Zwischenrouter gezwungen ist, Jumbo-Frames zu fragmentieren.

- NetApp empfiehlt* für ASA r2 SAN-Umgebungen die folgende Konfiguration:
- Jumbo-Frames sind bei 1GbE zwar wünschenswert, aber nicht erforderlich.
- Für maximale Leistung bei 10GbE und schnellerem iSCSI- und NVMe/TCP-Datenverkehr werden Jumbo-Frames benötigt.

TCP-Parameter

Drei Einstellungen sind oft falsch konfiguriert: TCP-Zeitstempel, selektive Bestätigung (SACK) und TCP-Fenster-Skalierung. Viele veraltete Dokumente im Internet empfehlen, einen oder mehrere dieser Parameter zu deaktivieren, um die Leistung zu verbessern. Vor vielen Jahren war diese Empfehlung verdienlich, als die CPU-Kapazitäten wesentlich geringer waren und der Overhead für die TCP-Verarbeitung, wenn möglich, reduziert werden konnte.

Bei modernen Betriebssystemen führt die Deaktivierung dieser TCP-Funktionen jedoch in der Regel nicht zu nachweisbaren Vorteilen und kann gleichzeitig die Leistung beeinträchtigen. In virtualisierten Netzwerkumgebungen sind Performance-Schäden besonders wahrscheinlich, da diese Funktionen für eine effiziente Handhabung von Paketverlusten und Änderungen der Netzwerkqualität erforderlich sind.

NetApp empfiehlt, TCP-Zeitstempel, SACK und TCP-Fenster-Skalierung auf dem Host zu aktivieren, und alle drei dieser Parameter sollten in jedem aktuellen Betriebssystem standardmäßig aktiviert sein.

FC SAN-Konfiguration

Bei der Konfiguration von FC SAN für Oracle-Datenbanken auf ASA r2-Systemen geht es in erster Linie darum, die gängigen Best Practices für SAN zu befolgen.

ASA r2 ist für reine SAN-Workloads optimiert, daher bleiben die Prinzipien die gleichen wie bei AFF/ FAS, wobei der Fokus auf Leistung, Ausfallsicherheit und Einfachheit liegt. Dies umfasst typische Planungsmaßnahmen wie die Sicherstellung einer ausreichenden Bandbreite im SAN zwischen Host und Speichersystem, die Überprüfung, ob alle SAN-Pfade zwischen allen erforderlichen Geräten vorhanden sind, die Verwendung der vom FC-Switch-Hersteller geforderten FC-Port-Einstellungen, die Vermeidung von ISL-Konflikten und die Verwendung einer ordnungsgemäßen SAN-Fabric-Überwachung.

Zoning

Eine FC-Zone sollte nie mehr als einen Initiator enthalten. Eine solche Anordnung mag zunächst zu funktionieren scheinen, doch Crosstalk zwischen Initiatoren beeinträchtigt letztendlich die Performance und Stabilität.

Multitarget-Zonen werden allgemein als sicher angesehen, obwohl in seltenen Fällen das Verhalten von FC-Zielports unterschiedlicher Anbieter Probleme verursacht hat. Es ist beispielsweise zu vermeiden, die Ziel-Ports von einem NetApp und einem nicht-NetApp Storage-Array in derselben Zone zu integrieren. Darüber hinaus besteht mit noch größerer Wahrscheinlichkeit die Gefahr, dass ein NetApp Storage-System und ein Bandgerät in dieselbe Zone platziert werden.

- ASA r2 verwendet Storage Availability Zones anstelle von Aggregaten, dies ändert jedoch nichts an den FC-Zonierungsprinzipien.
- Multipathing (MPIO) bleibt der primäre Ausfallsicherheitsmechanismus; allerdings sind bei ASA r2-Systemen, die symmetrisches Active-Active-Multipathing unterstützen, alle Pfade zu einer LUN aktiv und werden gleichzeitig für E/A genutzt.

Direct-Connect-Netzwerk

Storage-Administratoren ziehen es manchmal vor, ihre Infrastruktur zu vereinfachen, indem sie Netzwerk-Switches von der Konfiguration entfernen. Dies kann in einigen Szenarien unterstützt werden.

ISCSI und NVMe/TCP

Ein Host, der iSCSI oder NVMe/TCP verwendet, kann direkt an ein ASA r2-Speichersystem angeschlossen werden und normal funktionieren. Der Grund liegt in der Wegfindung. Direkte Verbindungen zu zwei verschiedenen Speicherkontrollern führen zu zwei unabhängigen Pfaden für den Datenfluss. Der Verlust eines Pfades, Ports oder Controllers verhindert nicht die Nutzung des anderen Pfades, vorausgesetzt, Multipathing ist korrekt konfiguriert.

FC-Direktverbindung

Es ist nicht möglich, einen Host über das FC-Protokoll direkt mit einem ASA r2-Speichersystem zu verbinden. Der Grund ist derselbe wie bei AFF/ FAS -Systemen: die Verwendung von NPIV. Der WWN, der einen ONTAP FC-Port im FC-Netzwerk identifiziert, verwendet eine Art von Virtualisierung namens NPIV. Jedes an ein ONTAP System angeschlossene Gerät muss in der Lage sein, eine NPIV-WWN zu erkennen. Derzeit gibt es keinen HBA-Anbieter, der einen HBA anbietet, der in einem Host installiert werden kann, der ein NPIV-Ziel unterstützen kann.

Storage-Konfiguration auf AFF/FAS Systemen

FC SAN

LUN-Ausrichtung

LUN-Ausrichtung bezieht sich auf die I/O-Optimierung in Bezug auf das zugrunde liegende Filesystem-Layout.

Auf einem ONTAP-System wird der Storage in 4-KB-Einheiten organisiert. Ein Datenbank- oder Filesystem-8-KB-Block sollte exakt zwei 4-KB-Blöcken zugeordnet werden. Wenn ein Fehler in der LUN-Konfiguration die Ausrichtung um 1 KB in beide Richtungen verschiebt, wäre jeder 8-KB-Block auf drei verschiedenen 4-KB-Storage-Blöcken vorhanden anstatt auf zwei. Diese Anordnung würde zu einer erhöhten Latenz führen und dazu führen, dass zusätzliche I/O-Vorgänge innerhalb des Speichersystems ausgeführt werden.

Die Ausrichtung wirkt sich auch auf LVM-Architekturen aus. Wenn ein physisches Volume innerhalb einer logischen Volume-Gruppe auf dem gesamten Laufwerk definiert wird (es werden keine Partitionen erstellt), wird der erste 4-KB-Block auf der LUN auf den ersten 4-KB-Block im Storage-System ausgerichtet. Dies ist eine korrekte Ausrichtung. Probleme ergeben sich bei Partitionen, da sie den Startort verschieben, an dem das Betriebssystem die LUN verwendet. Solange der Offset in ganzen 4-KB-Einheiten verschoben wird, ist die LUN ausgerichtet.

Erstellen Sie in Linux-Umgebungen logische Volume-Gruppen auf dem gesamten Laufwerkgerät. Wenn eine Partition erforderlich ist, überprüfen Sie die Ausrichtung, indem Sie ausführen `fdisk -u` und überprüfen, ob der Anfang jeder Partition ein Vielfaches von acht ist. Dies bedeutet, dass die Partition bei einem Vielfachen von acht 512-Byte-Sektoren beginnt, was 4 KB ist.

Siehe auch die Diskussion über die Ausrichtung der Kompressionsblöcke im Abschnitt "Effizienz". Jedes Layout, das an 8-KB-Komprimierungsblockgrenzen ausgerichtet ist, ist auch an 4-KB-Grenzen ausgerichtet.

Warnungen wegen Falschausrichtung

Die Datenbank-Wiederherstellungs-/Transaktionsprotokollierung erzeugt normalerweise nicht ausgerichtete I/O-Vorgänge, die irreführende Warnungen zu falsch ausgerichteten LUNs auf ONTAP verursachen können.

Die Protokollierung führt einen sequenziellen Schreibvorgang der Protokolldatei mit unterschiedlich großen Schreibvorgängen durch. Ein Protokollschreibvorgang, der sich nicht an 4-KB-Grenzen ausrichtet, verursacht

normalerweise keine Performance-Probleme, da der nächste Protokollschreibvorgang den Block abgeschlossen hat. Das Ergebnis: ONTAP ist in der Lage, fast alle Schreibvorgänge als komplette 4-KB-Blöcke zu verarbeiten, obwohl die Daten in einigen 4-KB-Blöcken in zwei separaten Operationen geschrieben wurden.

Überprüfen Sie die Ausrichtung mithilfe von Dienstprogrammen wie `sio` Oder `dd` Sie können I/O mit einer definierten Blockgröße generieren. Die I/O-Ausrichtungsstatistiken auf dem Storage-System können mit dem angezeigt werden `stats` Befehl. Siehe "Überprüfung der WAFL-Ausrichtung" Finden Sie weitere Informationen.

Die Ausrichtung in Solaris-Umgebungen ist komplizierter. Siehe "ONTAP SAN-Host-Konfiguration" Finden Sie weitere Informationen.

Achtung
Achten Sie in Solaris x86-Umgebungen besonders auf die richtige Ausrichtung, da die meisten Konfigurationen mehrere Ebenen von Partitionen haben. Solaris x86-Partitionsschichten befinden sich in der Regel oben auf einer Standard-Master-Bootdatensammelpartitionstabelle.

LUN-Dimensionierung und LUN-Anzahl

Die Auswahl der optimalen LUN-Größe und der Anzahl der zu verwendenden LUNs ist für optimale Performance und einfaches Management der Oracle-Datenbanken von entscheidender Bedeutung.

Eine LUN ist ein virtualisiertes Objekt auf ONTAP, das über alle Laufwerke im Hosting-Aggregat hinweg existiert. Die Performance der LUN wird daher von ihrer Größe nicht beeinflusst, da die LUN unabhängig von der gewählten Größe das volle Performance-Potenzial des Aggregats schöpft.

Aus praktischen Gründen möchten Kunden möglicherweise eine LUN einer bestimmten Größe verwenden. Wenn beispielsweise eine Datenbank auf einer LVM oder einer Oracle ASM-Datenträgergruppe erstellt wird, die aus zwei LUNs mit jeweils 1 TB besteht, muss diese Datenträgergruppe in Schritten von 1 TB erweitert werden. Es könnte besser sein, die Datenträgergruppe aus acht LUNs mit jeweils 500 GB zu erstellen, damit die Datenträgergruppe in kleineren Schritten erhöht werden kann.

Die Praxis, eine universelle Standard-LUN-Größe zu etablieren, wird davon abgeraten, da dies die Managebarkeit erschweren kann. Beispielsweise funktioniert eine standardmäßige LUN-Größe von 100 GB gut, wenn eine Datenbank oder ein Datastore im Bereich von 1 TB bis 2 TB liegt, jedoch erfordert eine Datenbank oder ein Datenspeicher mit einer Größe von 20 TB 200 LUNs. Das bedeutet, dass der Server-Neustart länger dauert, mehr Objekte in den verschiedenen Benutzeroberflächen zu verwalten sind und Produkte wie SnapCenter eine Erkennung für viele Objekte durchführen müssen. Derartige Probleme werden durch die Verwendung von weniger und größeren LUNs vermieden.

- Die Anzahl der LUNs ist wichtiger als die LUN-Größe.
- Die LUN-Größe wird überwiegend durch die Anforderungen der LUN-Anzahl gesteuert.
- Erstellen Sie nicht mehr LUNs als erforderlich.

LUN-Anzahl

Anders als die LUN-Größe wirkt sich die Anzahl der LUNs auf die Performance aus. Die Applikations-Performance hängt häufig von der Fähigkeit ab, parallelen I/O über die SCSI-Schicht auszuführen. Dadurch bieten zwei LUNs eine bessere Performance als eine einzelne LUN. Die Verwendung einer LVM wie Veritas VxVM, Linux LVM2 oder Oracle ASM ist die einfachste Methode, um die Parallelität zu erhöhen.

NetApp Kunden konnten im Allgemeinen nur einen minimalen Nutzen aus der Erhöhung der Anzahl von LUNs über sechzehn hinaus verzeichnen, obwohl sich bei den Tests mit 100 % SSD-Umgebungen mit sehr hoher zufälliger I/O-Last weitere Verbesserungen auf bis zu 64 LUNs gezeigt haben.

NetApp empfiehlt Folgendes:

Im Allgemeinen reichen vier bis sechzehn LUNs aus, um die I/O-Anforderungen jedes gegebenen Datenbank-Workloads zu unterstützen. Aufgrund der Einschränkungen bei Host-SCSI-Implementierungen könnten weniger als vier LUNs zu Performance-Einschränkungen führen.

LUN-Platzierung

Die optimale Platzierung von Datenbank-LUNs in ONTAP Volumes hängt in erster Linie davon ab, wie verschiedene ONTAP-Funktionen verwendet werden.

Volumes

Ein verbreiteter Verwechslungspunkt bei Kunden, die neu bei ONTAP sind, ist die Verwendung von FlexVols, die allgemein als „Volumes“ bezeichnet werden.

Ein Volume ist keine LUN. Diese Begriffe werden synonym mit vielen Produkten anderer Anbieter verwendet, darunter auch Cloud-Provider. ONTAP Volumes sind einfach Management-Container. Sie dienen nicht allein der Bereitstellung von Daten, noch belegen sie Speicherplatz. Sie sind Container für Dateien oder LUNs und sollen die Managebarkeit verbessern und vereinfachen, insbesondere bei großen Umgebungen.

Volumes und LUNs

Zugehörige LUNs befinden sich normalerweise in einem einzelnen Volume. Beispiel: Bei einer Datenbank, die 10 LUNs benötigt, sind normalerweise alle 10 LUNs auf demselben Volume platziert.

- Die Verwendung eines 1:1:1-Verhältnisses von LUNs zu Volumes, was einer LUN pro Volume entspricht, ist **nicht** eine formale Best Practice.
- Stattdessen sollten Volumes als Container für Workloads oder Datensätze angesehen werden. Es kann eine einzelne LUN pro Volume geben oder viele. Die richtige Antwort hängt von den Anforderungen an die Managebarkeit ab.
- Die Streuung von LUNs über eine unnötige Anzahl von Volumes kann zu zusätzlichem Overhead und Zeitplanungsproblemen bei Vorgängen wie Snapshot-Vorgängen führen, eine übermäßige Anzahl von Objekten, die in der UI angezeigt werden, und das Erreichen der Plattform-Volume-Grenzen führen, bevor das LUN-Limit erreicht wird.

Volumes, LUNs und Snapshots

Snapshot-Richtlinien und Zeitpläne werden auf dem Volume statt auf der LUN platziert. Ein Datensatz, der aus 10 LUNs besteht, würde nur eine einzige Snapshot-Politik erfordern, wenn diese LUNs auf demselben Volume Co-lokalisiert sind.

Darüber hinaus sorgt das Co-Lokalisieren aller verwandten LUNs für einen bestimmten Datensatz in einem einzelnen Volume für atomare Snapshot-Vorgänge. Beispielsweise könnte eine Datenbank, die auf 10 LUNs residierte, oder eine VMware-basierte Applikationsumgebung mit 10 verschiedenen Betriebssystemen als einzelnes, konsistentes Objekt gesichert werden, wenn alle zugrunde liegenden LUNs auf einem einzelnen Volume platziert werden. Wenn sie auf verschiedenen Volumes platziert werden, können die Snapshots zu

100% synchron sein, auch wenn sie zur gleichen Zeit geplant sind.

In manchen Fällen muss ein verwandter Satz von LUNs aufgrund von Recovery-Anforderungen in zwei verschiedene Volumes aufgeteilt werden. Beispielsweise könnte eine Datenbank vier LUNs für Datendateien und zwei LUNs für Protokolle haben. In diesem Fall könnte ein Datendatei-Volume mit 4 LUNs und ein Protokoll-Volume mit 2 LUNs die beste Option sein. Der Grund dafür ist eine unabhängige Wiederherstellbarkeit. Beispielsweise könnte das Datendatei-Volume selektiv in einen früheren Zustand zurückgesetzt werden. Dies bedeutet, dass alle vier LUNs auf den Status des Snapshot zurückgesetzt werden, während das Protokoll-Volume mit seinen kritischen Daten davon unberührt bleibt.

Volumes, LUNs und SnapMirror

SnapMirror Richtlinien und Operationen werden wie Snapshot-Vorgänge auf dem Volume, nicht auf der LUN durchgeführt.

Durch die Lokalisierung verwandter LUNs in einem einzelnen Volume können Sie eine einzelne SnapMirror Beziehung erstellen und alle enthaltenen Daten mit einem einzigen Update aktualisieren. Wie bei Snapshots wird auch das Update eine atomare Operation sein. Das SnapMirror Ziel würde garantiert über ein einzelnes Point-in-Time-Replikat der Quell-LUNs verfügen. Wenn die LUNs auf mehrere Volumes verteilt waren, können die Replikate miteinander konsistent sein oder nicht.

Volumes, LUNs und QoS

QoS kann selektiv auf einzelne LUNs angewendet werden, doch eine Festlegung auf Volume-Ebene ist in der Regel einfacher. So könnten beispielsweise alle LUNs, die die Gäste in einem bestimmten ESX Server nutzen, auf einem einzelnen Volume platziert werden, und anschließend könnte eine anpassungsfähige QoS-Richtlinie von ONTAP angewendet werden. Das Ergebnis ist ein selbst skalierendes Limit für IOPS pro TB, das für alle LUNs gilt.

Auch wenn eine Datenbank 100.000 IOPS benötigte und 10 LUNs belegte, wäre es einfacher, für ein einzelnes Volume eine einzige 100.000 IOPS-Grenze festzulegen, als 10 individuelle IOPS-Grenzwerte für 10.000 IOPS festzulegen, also eine für jede LUN.

Multi-Volume-Layouts

In einigen Fällen kann es von Vorteil sein, LUNs über mehrere Volumes zu verteilen. Der primäre Grund ist Controller-Striping. Ein HA-Storage-System kann beispielsweise eine einzige Datenbank hosten, für die das volle Verarbeitungs- und Caching-Potenzial jedes Controllers erforderlich ist. In diesem Fall würde ein typisches Design bedeuten, die Hälfte der LUNs in einem einzelnen Volume auf Controller 1 und die andere Hälfte der LUNs in einem einzelnen Volume auf Controller 2 zu platzieren.

Ebenso kann das Controller-Striping für den Lastausgleich verwendet werden. Ein HA-System, das 100 Datenbanken mit jeweils 10 LUNs hostet, kann so konzipiert werden, dass jede Datenbank auf jedem der beiden Controller ein 5-LUN-Volume erhält. Wenn zusätzliche Datenbanken bereitgestellt werden, wird die symmetrische Auslastung jedes Controllers gewährleistet.

Keines dieser Beispiele bezieht jedoch ein 1:1 Volume-zu-LUN-Verhältnis mit ein. Das Ziel bleibt die Optimierung des Managements durch Co-lokalisieren zugehöriger LUNs in Volumes.

Ein Verhältnis von 1:1 LUNs zu Volumes ist beispielsweise die Containerisierung, wobei jede LUN tatsächlich einen einzelnen Workload darstellt und individuell gemanagt werden muss. In solchen Fällen kann ein Verhältnis von 1:1 optimal sein.

LUN-Größe und LVM-Größe

Wenn ein SAN-basiertes Dateisystem seine Kapazitätsgrenze erreicht hat, gibt es zwei Möglichkeiten, den verfügbaren Speicherplatz zu erhöhen:

- Erhöhen Sie die Größe der LUNs
- Fügen Sie einer vorhandenen Volume-Gruppe eine LUN hinzu und vergrößern Sie das enthaltene logische Volume

Obwohl die LUN-Größenänderung eine Option ist, um die Kapazität zu erhöhen, ist es im Allgemeinen besser, eine LVM zu verwenden, einschließlich Oracle ASM. Einer der Hauptgründe für die Existenz von LVMs ist, dass keine LUN-Größe benötigt wird. Mit einer LVM werden mehrere LUNs zu einem virtuellen Speicherpool verknüpft. Die aus diesem Pool ausgearbeiteten logischen Volumes werden von der LVM gemanagt und können problemlos in der Größe geändert werden. Ein weiterer Vorteil besteht darin, dass Hotspots auf einem bestimmten Laufwerk vermieden werden, indem ein bestimmtes logisches Volume auf alle verfügbaren LUNs verteilt wird. Transparente Migration kann in der Regel mithilfe des Volume-Managers durchgeführt werden, um die zugrunde liegenden Extents eines logischen Volumes auf neue LUNs zu verschieben.

LVM-Striping

LVM-Striping bezieht sich auf die Verteilung von Daten über mehrere LUNs. So lässt sich die Performance vieler Datenbanken deutlich steigern.

Vor der Ära der Flash-Laufwerke wurde Striping verwendet, um die Performance-Einschränkungen rotierender Laufwerke zu überwinden. Beispiel: Wenn ein Betriebssystem einen Lesevorgang von 1 MB ausführen muss, würde das Lesen dieser 1 MB Daten von einem einzigen Laufwerk viel Festplattenkopf erfordern, der sucht und liest, da die 1 MB langsam übertragen wird. Wenn diese 1 MB Daten über 8 LUNs verteilt wurden, kann das Betriebssystem acht 128K-Lesevorgänge parallel ausführen und die für die 1-MB-Übertragung erforderliche Zeit verringern.

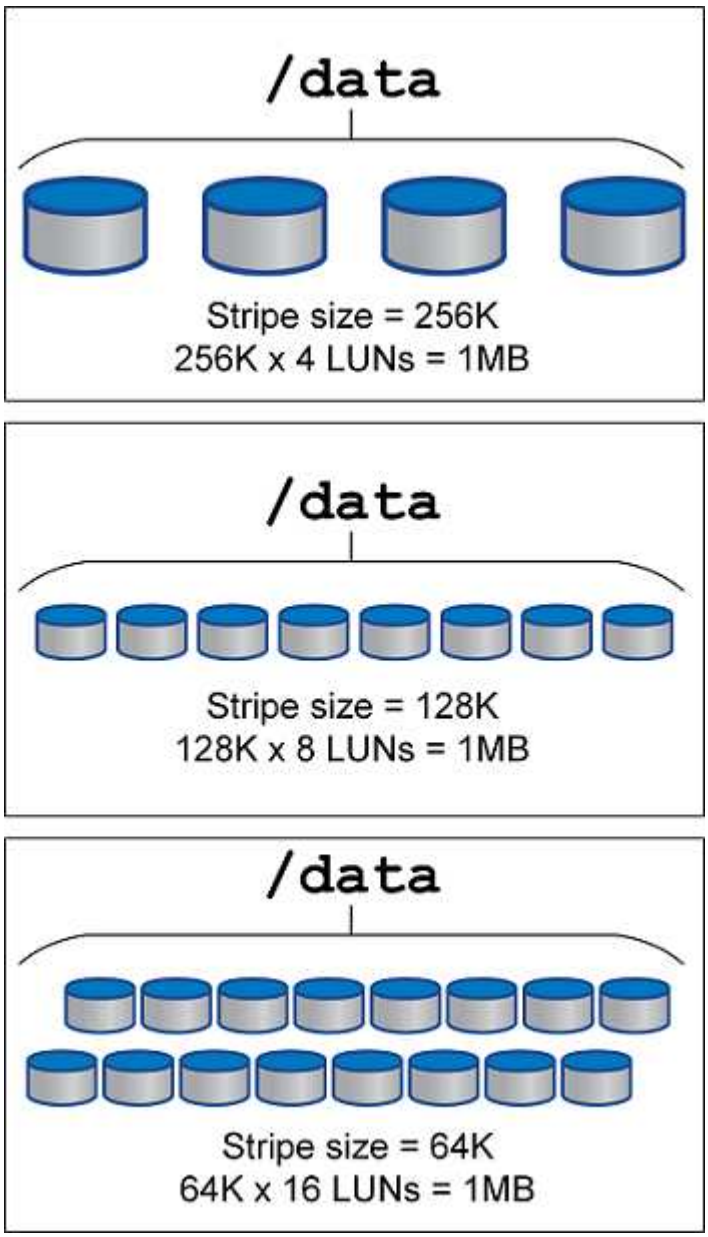
Das Striping mit rotierenden Laufwerken war schwieriger, da das I/O-Muster bereits im Vorfeld bekannt sein musste. Wenn das Striping nicht richtig auf die wahren I/O-Muster abgestimmt wurde, können Striping-Konfigurationen die Performance beeinträchtigen. Bei Oracle Datenbanken und insbesondere bei All-Flash-Konfigurationen ist Striping einfacher zu konfigurieren und hat sich nachweislich für eine drastische Verbesserung der Performance bewährt.

Logische Volume-Manager wie Oracle ASM Stripe sind standardmäßig aktiviert, aber native OS LVM nicht. Einige von ihnen verbinden mehrere LUNs als verkettete Geräte. Dies führt zu Datendateien, die auf einem und nur einem LUN-Gerät vorhanden sind. Dies verursacht Hotspots. Andere LVM-Implementierungen sind standardmäßig auf verteilte Extents eingestellt. Das ist ähnlich wie Striping, aber es ist gröber. Die LUNs in der Volume-Gruppe werden in große Teile geteilt, die als Extents bezeichnet werden und in der Regel in vielen Megabyte gemessen werden. Die logischen Volumes werden dann über diese Extents verteilt. Das Ergebnis ist ein zufälliger I/O-Vorgang für eine Datei, der auf LUNs verteilt werden sollte. Sequenzielle I/O-Vorgänge sind jedoch nicht so effizient wie möglich.

Die Performance-intensiven Applikations-I/O-Vorgänge erfolgen fast immer entweder (a) in Einheiten der grundlegenden Blockgröße oder (b) in Megabyte.

Das primäre Ziel einer Striped-Konfiguration ist es, sicherzustellen, dass Single-File I/O als eine Einheit ausgeführt werden kann. Multiblock-I/O, die eine Größe von 1 MB haben sollte, kann gleichmäßig über alle LUNs im Striped Volume hinweg parallelisiert werden. Das bedeutet, dass die Stripe-Größe nicht kleiner als die Blockgröße der Datenbank sein darf und die Stripe-Größe multipliziert mit der Anzahl der LUNs 1 MB betragen sollte.

Die folgende Abbildung zeigt drei mögliche Optionen für die Stripe-Größe und Breitenabstimmung. Die Anzahl der LUNs wird ausgewählt, um die oben beschriebenen Performance-Anforderungen zu erfüllen. In allen Fällen beträgt die Gesamtzahl der Daten innerhalb eines einzigen Stripes jedoch 1 MB.



NFS

Überblick

NetApp bietet seit über 30 Jahren NFS-Storage der Enterprise-Klasse. Seine Einsatzbereich wächst aufgrund der Einfachheit mit dem Trend zu Cloud-basierten Infrastrukturen.

Das NFS-Protokoll umfasst mehrere Versionen mit unterschiedlichen Anforderungen. Eine vollständige Beschreibung der NFS-Konfiguration mit ONTAP finden Sie unter "TR-4067 NFS on ONTAP Best Practices". In den folgenden Abschnitten werden einige der kritischeren Anforderungen und häufigen Benutzerfehler behandelt.

NFS-Versionen

Der NFS-Client des Betriebssystems muss von NetApp unterstützt werden.

- NFSv3 wird von Betriebssystemen unterstützt, die dem NFSv3 Standard folgen.
- NFSv3 wird vom Oracle dNFS-Client unterstützt.
- NFSv4 wird von allen Betriebssystemen unterstützt, die dem NFSv4-Standard entsprechen.
- Für NFSv4.1 und NFSv4.2 ist ein spezieller Support für das Betriebssystem erforderlich. Konsultieren Sie die "NetApp IMT" Für unterstützte Betriebssysteme.
- Oracle dNFS Unterstützung für NFSv4.1 erfordert Oracle 12.2.0.2 oder höher.

> Der "NetApp Support-Matrix" Für NFSv3 und NFSv4 sind keine spezifischen Betriebssysteme enthalten. Alle Betriebssysteme, die der RFC entsprechen, werden in der Regel unterstützt. Wenn Sie die Online-IMT nach Unterstützung für NFSv3 oder NFSv4 suchen, wählen Sie kein bestimmtes Betriebssystem aus, da keine Treffer angezeigt werden. Alle Betriebssysteme werden implizit von der allgemeinen Richtlinie unterstützt.

Linux NFSv3 TCP-Slot-Tabellen

TCP-Slot-Tabellen sind das NFSv3 Äquivalent zur Warteschlangentiefe des Host Bus Adapters (HBA). Diese Tabellen steuern die Anzahl der NFS-Vorgänge, die zu einem beliebigen Zeitpunkt ausstehen können. Der Standardwert ist normalerweise 16, was für eine optimale Performance viel zu niedrig ist. Das entgegengesetzte Problem tritt auf neueren Linux-Kerneln auf, die automatisch die Begrenzung der TCP-Slot-Tabelle auf ein Niveau erhöhen können, das den NFS-Server mit Anforderungen sättigt.

Um eine optimale Performance zu erzielen und Performance-Probleme zu vermeiden, passen Sie die Kernel-Parameter an, die die TCP-Slot-Tabellen steuern.

Führen Sie die aus `sysctl -a | grep tcp.*.slot_table` Und beobachten Sie die folgenden Parameter:

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

Alle Linux-Systeme sollten enthalten `sunrpc.tcp_slot_table_entries`, Aber nur einige enthalten `sunrpc.tcp_max_slot_table_entries`. Beide sollten auf 128 gesetzt werden.

> Wenn diese Parameter nicht eingestellt werden, kann dies erhebliche Auswirkungen auf die Leistung haben. In einigen Fällen ist die Performance eingeschränkt, da das linux-Betriebssystem nicht genügend I/O ausgibt In anderen Fällen erhöht sich die I/O-Latenz, wenn das linux Betriebssystem versucht, mehr I/O-Vorgänge auszustellen, als gewartet werden kann.

AdR und NFS

Einige Kunden haben Performance-Probleme gemeldet, die auf übermäßig viele I/O-Vorgänge für Daten im führen `ADR` Standort. Das Problem tritt in der Regel erst auf, wenn sich viele Performance-Daten angesammelt haben. Der Grund für den übermäßigen I/O ist unbekannt, aber dieses Problem scheint darauf zurückzuführen zu sein, dass Oracle-Prozesse das Zielverzeichnis wiederholt auf Änderungen scannen.

Entfernen des `noac` Und/oder `actimeo=0` Mount-Optionen ermöglichen das Caching des Host-Betriebssystems und reduzieren die Storage-I/O-Level.

NetApp empfiehlt nicht zu platzieren `ADR` Daten auf einem Filesystem mit `noac` Oder `actimeo=0` Weil Performance-Probleme wahrscheinlich sind. Trennen `ADR` Daten an einen anderen Bereitstellungspunkt, falls erforderlich.

Nur nfs-Rootonly und Mount-Rootonly

ONTAP enthält die NFS-Option `nfs-rootonly` Damit wird gesteuert, ob der Server NFS-Datenverkehrsverbindungen von hohen Ports akzeptiert. Als Sicherheitsmaßnahme ist es nur dem Root-Benutzer erlaubt, TCP/IP-Verbindungen über einen Quellport unter 1024 zu öffnen, da solche Ports normalerweise für die Verwendung durch das Betriebssystem und nicht für Benutzerprozesse reserviert sind. Durch diese Einschränkung wird sichergestellt, dass NFS-Datenverkehr von einem tatsächlichen Betriebssystem-NFS-Client stammt und kein schädlicher Prozess, der einen NFS-Client emuliert. Der Oracle dNFS-Client ist ein Benutzerspeichertreiber, aber der Prozess läuft als root, daher ist es in der Regel nicht erforderlich, den Wert von zu ändern `nfs-rootonly`. Die Verbindungen werden von niedrigen Ports hergestellt.

Der `mount-rootonly` Die Option gilt nur für NFSv3. Er steuert, ob der RPC-MOUNT-Aufruf von Ports über 1024 akzeptiert wird. Wenn dNFS verwendet wird, läuft der Client wieder als root, so dass er Ports unter 1024 öffnen kann. Dieser Parameter hat keine Auswirkung.

Prozesse, die Verbindungen mit dNFS über NFS Version 4.0 und höher öffnen, laufen nicht als Root und erfordern daher Ports über 1024. Der `nfs-rootonly` Der Parameter muss auf disabled gesetzt werden, damit dNFS die Verbindung herstellen kann.

Wenn `nfs-rootonly` Ist aktiviert, ist das Ergebnis ein Hängezustand während der Mount-Phase beim Öffnen von dNFS-Verbindungen. Der sqlplus-Ausgang sieht ähnlich aus wie folgt:

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

Der Parameter kann wie folgt geändert werden:

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

In seltenen Fällen müssen Sie möglicherweise sowohl nfs-rootonly als auch Mount-rootonly auf disabled ändern. Wenn ein Server eine extrem große Anzahl von TCP-Verbindungen verwaltet, ist es möglich, dass keine Ports unter 1024 verfügbar sind und das Betriebssystem gezwungen ist, höhere Ports zu verwenden. Diese beiden ONTAP-Parameter müssen geändert werden, damit die Verbindung abgeschlossen werden kann.

NFS-Export-Richtlinien: Superuser und setuid

Wenn sich Oracle-Binärdateien auf einer NFS-Freigabe befinden, muss die Exportrichtlinie Superuser- und setuid-Berechtigungen enthalten.

Für allgemeine Fileservices wie Home Directories der Benutzer verwendete Shared NFS-Exporte vernichten normalerweise den Root-Benutzer. Dies bedeutet, dass eine Anfrage des Root-Benutzers auf einem Host, der ein Dateisystem gemountet hat, als anderer Benutzer mit niedrigeren Berechtigungen neu zugeordnet wird. Dies hilft, Daten zu sichern, indem ein Root-Benutzer auf einem bestimmten Server daran gehindert wird, auf Daten auf dem freigegebenen Server zuzugreifen. Das setuid-Bit kann auch ein Sicherheitsrisiko in einer gemeinsam genutzten Umgebung darstellen. Mit dem setuid-Bit kann ein Prozess als ein anderer Benutzer ausgeführt werden als der Benutzer, der den Befehl aufruft. Beispielsweise wird ein Shell-Skript, das im Besitz von root war, mit dem setuid-Bit als root ausgeführt. Wenn dieses Shell-Skript von anderen Benutzern geändert werden könnte, könnte jeder Benutzer, der nicht root ist, einen Befehl als root ausgeben, indem er das Skript aktualisiert.

Die Oracle-Binärdateien enthalten Dateien im Besitz von root und verwenden das setuid-Bit. Wenn Oracle-Binärdateien auf einer NFS-Freigabe installiert sind, muss die Exportrichtlinie die entsprechenden Superuser- und setuid-Berechtigungen enthalten. Im folgenden Beispiel enthält die Regel beides `allow-suid` Und Genehmigungen `superuser` (Root)-Zugriff für NFS-Clients unter Verwendung der Systemauthentifizierung.

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

Konfiguration von NFSv4/4.1

Für die meisten Applikationen gibt es kaum einen Unterschied zwischen NFSv3 und NFSv4. Applikations-I/O ist in der Regel sehr einfach I/O und nicht von einigen der erweiterten Funktionen, die in NFSv4 verfügbar sind, erheblich profitieren. Höhere Versionen von NFS sollten nicht aus Sicht des Datenbank-Storage als „Upgrade“ betrachtet werden, sondern als Versionen von NFS, die zusätzliche Features enthalten. Wenn beispielsweise die End-to-End-Sicherheit des kerberos Datenschutzmodus (krb5p) erforderlich ist, ist NFSv4 erforderlich.

NetApp empfiehlt NFSv4.1 zu verwenden, wenn NFSv4-Funktionen erforderlich sind. Es gibt einige funktionale Verbesserungen am NFSv4-Protokoll in NFSv4.1, die die Ausfallsicherheit in bestimmten Edge-Fällen verbessern.

Der Wechsel zu NFSv4 ist komplizierter als einfach die Mount-Optionen von vers=3 auf vers=4.1 zu ändern. Eine ausführlichere Erläuterung der NFSv4-Konfiguration mit ONTAP, einschließlich Anleitungen zur Konfiguration des Betriebssystems, finden Sie unter "TR-4067 NFS on ONTAP Best Practices". Die folgenden Abschnitte dieses TR erklären einige der Grundvoraussetzungen für die Verwendung von NFSv4.

NFSv4-Domäne

Eine vollständige Erklärung der NFSv4/4.1-Konfiguration geht über den Umfang dieses Dokuments hinaus, aber ein häufig aufgetretendes Problem ist eine Diskrepanz bei der Domänenzuordnung. Aus Sicht von sysadmin scheinen sich die NFS-Dateisysteme normal zu verhalten, aber Anwendungen melden Fehler über Berechtigungen und/oder setuid auf bestimmte Dateien. In einigen Fällen haben Administratoren fälschlicherweise festgestellt, dass die Berechtigungen der Anwendungsbinärdateien beschädigt wurden und chown- oder chmod-Befehle ausgeführt haben, wenn das eigentliche Problem der Domänenname war.

Der NFSv4-Domänenname wird auf der ONTAP SVM festgelegt:

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

Der NFSv4-Domänenname auf dem Host wird in festgelegt `/etc/idmap.cfg`

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

Die Domänennamen müssen übereinstimmen. Wenn dies nicht der Fall ist, werden ähnliche Zuordnungsfehler wie die folgenden in angezeigt `/var/log/messages`:

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

Anwendungsbinärdateien, wie z. B. Oracle-Datenbank-Binärdateien, enthalten Dateien im Besitz von root mit dem setuid-Bit, was bedeutet, dass eine Diskrepanz in den NFSv4-Domänennamen Fehler beim Starten von Oracle verursacht und eine Warnung über die Eigentumsrechte oder Berechtigungen einer Datei namens enthält `oradism`, Die sich im befindet `$ORACLE_HOME/bin` Verzeichnis. Sie sollte wie folgt aussehen:

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

Wenn diese Datei mit der Eigentümerschaft von Niemand angezeigt wird, kann es ein Problem mit der NFSv4-Domänenzuordnung geben.

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

Um dies zu beheben, überprüfen Sie die `/etc/idmap.cfg` Datei mit der v4-id-Domain-Einstellung auf ONTAP und stellen Sie sicher, dass sie konsistent sind. Wenn dies nicht der Fall ist, nehmen Sie die erforderlichen Änderungen vor, und führen Sie aus `nfsidmap -c`, Und warten Sie einen Moment, bis sich die Änderungen fortpflanzen. Die Dateieigentümerschaft sollte dann ordnungsgemäß als root erkannt werden. Wenn ein Benutzer versucht hatte, ausgeführt zu werden `chown root` Vor der Korrektur der Konfiguration der NFS-Domänen in dieser Datei muss möglicherweise ausgeführt werden `chown root` Ein weiteres Jahr in der

Oracle Direct NFS (dNFS)

Oracle Databases können NFS auf zweierlei Weise verwenden.

Zunächst kann es ein Dateisystem verwenden, das mit dem nativen NFS-Client gemountet ist, der Teil des Betriebssystems ist. Dies wird manchmal Kernel NFS oder kNFS genannt. Das NFS-Dateisystem ist gemountet und von der Oracle-Datenbank genau so verwendet wie jede andere Anwendung ein NFS-Dateisystem verwenden würde.

Die zweite Methode ist Oracle Direct NFS (dNFS). Hierbei handelt es sich um eine Implementierung des NFS-Standards in der Oracle Datenbanksoftware. Die Art und Weise, wie Oracle-Datenbanken vom DBA konfiguriert oder verwaltet werden, bleibt unverändert. Sofern das Storage-System selbst die richtigen Einstellungen hat, sollte die Verwendung von dNFS für das DBA-Team und die Endanwender transparent sein.

Eine Datenbank mit aktivierter dNFS-Funktion hat noch die üblichen NFS-Dateisysteme gemountet. Sobald die Datenbank geöffnet ist, öffnet die Oracle-Datenbank eine Reihe von TCP/IP-Sitzungen und führt NFS-Vorgänge direkt aus.

Direktes NFS

Der Hauptwert von Direct NFS von Oracle besteht darin, den NFS-Client des Hosts zu umgehen und NFS-Dateivorgänge direkt auf einem NFS-Server auszuführen. Wenn Sie diese Option aktivieren, muss nur die Oracle Disk Manager (ODM)-Bibliothek geändert werden. Anweisungen zu diesem Prozess finden Sie in der Oracle-Dokumentation.

Die Verwendung von dNFS führt zu einer deutlichen Verbesserung der I/O-Performance und verringert die Last auf dem Host und dem Storage-System, da I/O so effizient wie möglich ausgeführt wird.

Darüber hinaus enthält Oracle dNFS eine **Option** für Multipathing und Fehlertoleranz der Netzwerkschnittstelle. Beispielsweise können zwei 10-GB-Schnittstellen verbunden werden, um eine Bandbreite von 20 GB bereitzustellen. Ein Ausfall einer Schnittstelle führt dazu, dass die I/O-Vorgänge auf der anderen Schnittstelle wiederholt werden. Der gesamte Vorgang ähnelt dem FC-Multipathing. Multipathing war schon vor Jahren üblich, als 1 GB ethernet der häufigste Standard war. Für die meisten Oracle Workloads ist eine 10-Gbit-NIC ausreichend. Wird jedoch mehr benötigt, können 10-Gbit-NICs verbunden werden.

Wenn dNFS verwendet wird, ist es wichtig, dass alle Patches, die in Oracle Doc 1495104.1 beschrieben werden, installiert sind. Wenn ein Patch nicht installiert werden kann, muss die Umgebung überprüft werden, um sicherzustellen, dass die in diesem Dokument beschriebenen Fehler keine Probleme verursachen. In manchen Fällen kann dNFS nicht verwendet werden, da die erforderlichen Patches nicht installiert werden können.

Verwenden Sie dNFS nicht mit Round-Robin-Namensauflösungen wie DNS, DDNS, NIS oder anderen Methoden. Dazu gehört auch die in ONTAP verfügbare DNS-Lastausgleichsfunktion. Wenn eine Oracle-Datenbank mit dNFS einen Hostnamen in eine IP-Adresse auflöst, darf sie sich bei nachfolgenden Suchen nicht ändern. Dies kann zu Abstürzen der Oracle-Datenbank und einer möglichen Beschädigung von Daten führen.

Aktivieren von dNFS

Oracle dNFS kann mit NFSv3 ohne Konfiguration arbeiten, die über die Aktivierung der dNFS Library hinaus erforderlich ist (siehe Oracle Dokumentation für den spezifischen Befehl erforderlich), aber wenn dNFS keine Verbindung herstellen kann, kann es im Hintergrund zurück zum Kernel NFS Client zurückkehren. In einem solchen Fall kann die Performance erheblich beeinträchtigt werden.

Wenn Sie dNFS-Multiplexing über mehrere Schnittstellen, mit NFSv4.X, oder Verschlüsselung verwenden

möchten, müssen Sie eine oranfstab-Datei konfigurieren. Die Syntax ist extrem streng. Kleine Fehler in der Datei können dazu führen, dass der Start hängend oder umgangen die oranfstab-Datei.

Zum Zeitpunkt der Erstellung dieses Berichts funktioniert dNFS-Multipathing nicht mit NFSv4.1 und aktuellen Versionen von Oracle Database. Eine oranfstab-Datei, die NFSv4.1 als Protokoll angibt, kann nur eine Single Path-Anweisung für einen bestimmten Export verwenden. Der Grund dafür ist, dass ONTAP Client ID Trunking nicht unterstützt. Oracle Database Patches zur Behebung dieser Einschränkung sind möglicherweise in Zukunft verfügbar.

Der einzige Weg, um sicher zu sein, dNFS funktioniert wie erwartet, ist die Abfrage der V-dnfs-Tabellen.

Unten finden Sie ein Beispiel für eine Oranfstab-Datei unter /etc. Dies ist einer von mehreren Speicherorten, an denen eine oranfstab-Datei platziert werden kann.

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

Im ersten Schritt wird überprüft, ob dNFS für die angegebenen Dateisysteme betriebsbereit ist:

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

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

/dbf
NFSv3.0
```

Diese Ausgabe zeigt an, dass dNFS mit diesen beiden Dateisystemen verwendet wird, aber es bedeutet **Not**, dass oranfstab betriebsbereit ist. Wenn ein Fehler aufgetreten ist, hätte dNFS die NFS-Dateisysteme des Hosts automatisch erkannt und Sie können immer noch die gleiche Ausgabe von diesem Befehl sehen.

Multipathing kann wie folgt überprüft werden:

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

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

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

NFSv3test
jfs_svmdr-nfs2
         1

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

NFSv3test
jfs_svmdr-nfs1
         0

NFSv3test
jfs_svmdr-nfs2

[output truncated]

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

NFSv3test
jfs_svmdr-nfs1
         0

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

```
----------

NFSv3test
jfs_svmdr-nfs2
         1


66 rows selected.
```

Das sind die Verbindungen, die dNFS verwendet. Für jeden SVRNAME-Eintrag sind zwei Pfade und Kanäle sichtbar. Das bedeutet, dass Multipathing funktioniert, was bedeutet, dass die oranfstab-Datei erkannt und verarbeitet wurde.

Direkter NFS- und Host-Filesystem-Zugriff

Die Verwendung von dNFS kann gelegentlich Probleme für Applikationen oder Benutzeraktivitäten verursachen, die auf den sichtbaren Filesystemen basieren, die auf dem Host gemountet sind, da der dNFS-Client vom Host-Betriebssystem aus auf das Filesystem zugreift. Der dNFS-Client kann Dateien ohne Kenntnis des Betriebssystems erstellen, löschen und ändern.

Wenn die Mount-Optionen für Single-Instance-Datenbanken verwendet werden, ermöglichen sie das Caching von Datei- und Verzeichnisattributen, was auch bedeutet, dass der Inhalt eines Verzeichnisses zwischengespeichert wird. Daher kann dNFS eine Datei erstellen, und es gibt eine kurze Verzögerung, bevor das Betriebssystem den Verzeichnisinhalt erneut liest und die Datei für den Benutzer sichtbar wird. Dies ist in der Regel kein Problem, aber in seltenen Fällen können Dienstprogramme wie SAP BR*Tools Probleme haben. Beheben Sie in diesem Fall das Problem, indem Sie die Mount-Optionen ändern, um die Empfehlungen für Oracle RAC zu verwenden. Mit dieser Änderung wird das gesamte Host-Caching deaktiviert.

Mount-Optionen nur ändern, wenn (a) dNFS verwendet wird und (b) ein Problem auf eine Verzögerung bei der Dateisichtbarkeit zurückzuführen ist. Wenn dNFS nicht verwendet wird, führt die Verwendung der Oracle RAC Mount-Optionen auf einer Single-Instance-Datenbank zu einer verminderte Performance.

> In der Anmerkung zu `nosharecache` in "Mount-Optionen für Linux NFS" finden Sie ein Linux-spezifisches dNFS-Problem, das zu ungewöhnlichen Ergebnissen führen kann.

NFS-Lease und -Sperren

NFSv3 ist statusfrei. Das bedeutet effektiv, dass der NFS-Server (ONTAP) nicht verfolgt, welche Dateisysteme gemountet sind, von wem oder welche Sperren tatsächlich vorhanden sind.

ONTAP verfügt über einige Funktionen, die Mount-Versuche aufzeichnen, sodass Sie eine Vorstellung davon haben, welche Clients möglicherweise auf Daten zugreifen, und es gibt möglicherweise Hinweissperren, aber diese Informationen sind nicht garantiert zu 100% vollständig. Es kann nicht vollständig sein, da die Nachverfolgung des NFS-Client-Status nicht Teil des NFSv3-Standards ist.

Status der NFSv4-Daten

Im Gegensatz dazu ist NFSv4 zustandsbehafteten. Der NFSv4-Server verfolgt, welche Clients welche Dateisysteme verwenden, welche Dateien existieren, welche Dateien und/oder Regionen von Dateien gesperrt sind usw. Dies bedeutet, dass eine regelmäßige Kommunikation zwischen einem NFSv4-Server erforderlich

ist, um die Statusdaten auf dem aktuellen Stand zu halten.

Die wichtigsten Zustände, die vom NFS-Server verwaltet werden, sind NFSv4-Locks und NFSv4-Leases, und sie sind sehr miteinander verflochten. Sie müssen verstehen, wie jede einzelne von sich aus funktioniert und wie sie miteinander in Beziehung stehen.

NFSv4-Sperren

Bei NFSv3 sind Sperren Empfehlung. Ein NFS-Client kann weiterhin „gesperrte“ Dateien ändern oder löschen. Eine NFSv3-Sperre läuft nicht von selbst ab, sie muss entfernt werden. Dies führt zu Problemen. Wenn Sie beispielsweise über eine geclusterte Applikation verfügen, die NFSv3-Sperren erstellt, und einer der Nodes ausfällt, wie gehen Sie vor? Sie können die Anwendung auf den verbleibenden Knoten codieren, um die Sperren zu entfernen, aber wie können Sie wissen, dass das sicher ist? Vielleicht ist der „ausgefallene“ Knoten funktionsfähig, kommuniziert aber nicht mit dem Rest des Clusters?

Mit NFSv4 haben Sperren eine begrenzte Dauer. Solange der Client mit den Sperren weiterhin mit dem NFSv4-Server eincheckt, darf kein anderer Client diese Sperren erwerben. Wenn ein Client nicht mit dem NFSv4 eincheckt, werden die Sperren schließlich vom Server widerrufen und andere Clients können Sperren anfordern und erhalten.

NFSv4-Leasing

NFSv4-Sperren sind einem NFSv4-Leasing zugeordnet. Wenn ein NFSv4-Client eine Verbindung mit einem NFSv4-Server herstellt, erhält er eine Leasing-Option. Wenn der Kunde eine Sperre erhält (es gibt viele Arten von Sperren), dann ist die Sperre mit dem Leasing verbunden.

Diese Lease hat ein definiertes Timeout. Standardmäßig setzt ONTAP den Timeout-Wert auf 30 Sekunden:

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

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

Dies bedeutet, dass ein NFSv4-Client alle 30 Sekunden mit dem NFSv4-Server einchecken muss, um seine Mietverträge zu erneuern.

Der Lease wird automatisch durch jede Aktivität erneuert, sodass, wenn der Kunde arbeitet, keine zusätzlichen Operationen durchgeführt werden müssen. Wenn eine Anwendung still wird und keine echte Arbeit macht, muss sie stattdessen eine Art Keep-Alive-Vorgang (SEQUENZ genannt) durchführen. Es ist im Grunde nur sagen: "Ich bin immer noch hier, bitte aktualisieren Sie meine Mietverträge."

```
 *Question:* What happens if you lose network connectivity for 31 seconds?
NFSv3 ist statusfrei. Es wird keine Kommunikation der Clients erwartet.
NFSv4 ist zustandsbehaftet. Sobald dieser Leasingzeitraum verstrichen ist,
läuft der Leasingvertrag ab, Sperren werden aufgehoben und die gesperrten
Dateien werden anderen Clients zur Verfügung gestellt.
```

Mit NFSv3 können Sie Netzwerkkabel umlegen, Netzwerk-Switches neu booten, Konfigurationsänderungen vornehmen und ziemlich sicher sein, dass nichts Schlimmes passiert. Anwendungen würden normalerweise

nur geduldig warten, bis die Netzwerkverbindung wieder funktioniert.

Mit NFSv4 haben Sie 30 Sekunden (es sei denn, Sie haben den Wert dieses Parameters innerhalb von ONTAP erhöht), um Ihre Arbeit abzuschließen. Wenn Sie das überschreiten, Ihre Leasing-Zeit aus. Normalerweise führt dies zu einem Absturz der Anwendung.

Wenn Sie beispielsweise über eine Oracle-Datenbank verfügen und die Netzwerkverbindung (manchmal auch als „Netzwerkpartition“ bezeichnet) unterbrochen wird, die das Lease-Timeout überschreitet, stürzt die Datenbank ab.

Dies ist ein Beispiel dafür, was im Oracle-Alarmprotokoll passiert, wenn dies geschieht:

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

Wenn Sie sich die Syslogs ansehen, sollten Sie mehrere der folgenden Fehler sehen:

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

Die Protokollmeldungen sind in der Regel das erste Anzeichen eines Problems, das nicht durch das Einfrieren der Anwendung verursacht wird. In der Regel sehen Sie während des Netzwerkausfalls überhaupt nichts, da Prozesse und das Betriebssystem selbst blockiert sind und versuchen, auf das NFS-Dateisystem zuzugreifen.

Die Fehler werden angezeigt, nachdem das Netzwerk wieder betriebsbereit ist. Im obigen Beispiel hat das Betriebssystem versucht, die Sperren nach der Wiederherstellung der Verbindung erneut zu erfassen, aber es war zu spät. Der Mietvertrag war abgelaufen und die Schlösser wurden entfernt. Dies führt zu einem Fehler, der sich auf die Oracle-Ebene ausbreitet und die Meldung im Alarmprotokoll verursacht. Je nach Version und Konfiguration der Datenbank können Sie Abweichungen von diesen Mustern sehen.

Zusammenfassend lässt sich sagen, dass NFSv3 eine Netzwerkunterbrechung toleriert, aber NFSv4 ist sensibler und sieht einen definierten Leasing-Zeitraum vor.

Was ist, wenn eine 30-Sekunden-Zeitüberschreitung nicht akzeptabel ist? Was tun Sie, wenn Sie ein dynamisch verändertes Netzwerk verwalten, in dem Switches neu gestartet oder Kabel verlegt werden, und das Ergebnis ist eine gelegentliche Netzwerkunterbrechung? Sie könnten die Leasingdauer verlängern, aber ob Sie dies tun möchten, erfordert eine Erklärung der NFSv4-Kulanzzeiträume.

NFSv4-Kulanzzeiträume

Wenn ein NFSv3 Server neu gestartet wird, ist er fast sofort in der Lage, I/O zu bedienen. Es war nicht die Aufrechterhaltung einer Art von Zustand über Kunden. Dies führt dazu, dass ein ONTAP-Übernahmevorgang oft fast unmittelbar zu erfolgen scheint. Sobald ein Controller bereit ist, mit der Datenbereitstellung zu beginnen, sendet er ein ARP an das Netzwerk, das die Änderung der Topologie signalisiert. Clients erkennen dies normalerweise nahezu sofort, und die Daten werden wieder fließend gespeichert.

NFSv4 erzeugt jedoch eine kurze Pause. Nur ein Teil davon, wie NFSv4 funktioniert.

> Die folgenden Abschnitte sind aktuell ab ONTAP 9.15.1, aber das Lease- und Sperrverhalten sowie Tuning-Optionen können von Version zu Version wechseln. Wenn Sie NFSv4-Leasing/Lock-Timeouts einstellen müssen, konsultieren Sie bitte den NetApp-Support für die neuesten Informationen.

NFSv4-Server müssen die Leasing-Optionen, Sperren und die Verwendung welcher Daten verfolgen. Wenn ein NFS-Server in Panik Gerät und neu startet oder einen Moment lang Strom verliert oder während der Wartungsaktivitäten neu gestartet wird, führt dies zu einer Lease/Sperre und zum Verlust anderer Clientinformationen. Der Server muss herausfinden, welcher Client welche Daten verwendet, bevor er den Betrieb wiederaufnehmen kann. Hier kommt die Kulanzzeit ins Spiel.

Wenn Sie Ihren NFSv4-Server plötzlich aus- und wieder einschalten. Wenn es wieder verfügbar ist, erhalten Kunden, die versuchen, die E/A-Vorgänge fortzusetzen, eine Antwort, die im Wesentlichen besagt: „Ich habe die Leasing-/Sperrdaten verloren. Möchten Sie Ihre Sperren erneut registrieren?“ Das ist der Anfang der Gnadenfrist. Die Standardeinstellung ist 45 Sekunden bei ONTAP:

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

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

Das Ergebnis ist, dass ein Controller nach einem Neustart I/O-Vorgänge pausiert, während alle Clients ihre Mietverträge und Sperren zurückfordern. Nach Ablauf der Kulanzzeit nimmt der Server die E/A-Vorgänge wieder auf.

Diese Kulanzzeit steuert die Rückgewinnung von Leasing-Verträgen während Änderungen an der Netzwerkschnittstelle, aber es gibt eine zweite Kulanzzeit, die die Rückgewinnung während des Speicher-Failovers steuert `locking.grace_lease_seconds`. Hierbei handelt es sich um eine Option auf Node-Ebene.

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

Wenn Sie beispielsweise häufig LIF-Failovers durchführen mussten und die Gnadenfrist reduzieren mussten,

würden Sie ändern `v4-grace-seconds`. Wenn Sie die IO Wiederaufnahme Zeit während des Controller-Failovers verbessern wollten, müssten Sie ändern `locking.grace_lease_seconds`.

Ändern Sie diese Werte nur mit Vorsicht und nach vollständiger Kenntnis der Risiken und Konsequenzen. Die I/O-Pausen, die mit Failover- und Migrationsvorgängen mit NFSv4.X verbunden sind, können nicht vollständig vermieden werden. Sperrfristen, Lease- und Kulanzfristen sind Teil der NFS RFC. Für viele Kunden ist NFSv3 vorzuziehen, da Failover-Zeiten schneller sind.

Leasing-Timeouts im Vergleich zu Kulanzzeiträumen

Die Kulanzzeit und die Leasingdauer sind miteinander verknüpft. Wie bereits erwähnt, beträgt das standardmäßige Leasingzeitlimit 30 Sekunden, was bedeutet, dass NFSv4-Clients mindestens alle 30 Sekunden beim Server einchecken müssen, oder sie verlieren ihre Leasingverhältnisse und damit ihre Sperren. Die Kulanzzeit ist vorhanden, um einem NFS-Server zu ermöglichen, Lease/Lock-Daten neu zu erstellen, und es ist standardmäßig 45 Sekunden. Die Kulanzzeit muss länger als die Leasingfrist sein. Dadurch wird sichergestellt, dass eine NFS-Client-Umgebung, die zur Verlängerung von Leasingverträgen mindestens alle 30 Sekunden entwickelt wurde, nach einem Neustart beim Server einchecken kann. Eine Nachfrist von 45 Sekunden sorgt dafür, dass alle Kunden, die erwarten, ihre Mietverträge mindestens alle 30 Sekunden auf jeden Fall die Möglichkeit haben, dies zu tun.

Wenn ein Timeout von 30 Sekunden nicht akzeptabel ist, können Sie die Leasingdauer verlängern.

Wenn Sie das Lease-Timeout auf 60 Sekunden erhöhen möchten, um einem Netzwerkausfall von 60 Sekunden standzuhalten, müssen Sie auch die Kulanzzeit verlängern. Das bedeutet, dass Sie längere I/O-Pausen während Controller-Failover erleben.

Das sollte normalerweise kein Problem sein. In der Regel aktualisieren ONTAP Controller nur ein oder zwei Mal pro Jahr, und ein ungeplanter Failover aufgrund von Hardwareausfällen ist äußerst selten. Darüber hinaus würden Sie bei einem Netzwerk, wo ein Netzwerkausfall von 60 Sekunden zu besorgen war und Sie eine Leasingzeit von 60 Sekunden benötigen, wahrscheinlich auch keinem seltenen Storage-System-Failover widersprechen, was zu einer Pause von 61 Sekunden führt. Sie haben bereits bestätigt, dass Sie ein Netzwerk haben, das ziemlich häufig über 60 Sekunden anhält.

NFS-Caching

Das Vorhandensein einer der folgenden Mount-Optionen bewirkt, dass das Host-Caching deaktiviert wird:

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

Diese Einstellungen können sich stark negativ auf die Geschwindigkeit der Softwareinstallation, des Patches und der Backup-/Wiederherstellungsvorgänge auswirken. In manchen Fällen, insbesondere bei geclusterten Applikationen, sind diese Optionen unweigerlich erforderlich, weil die Cache-Kohärenz über alle Nodes im Cluster hinweg gewährleistet werden muss. In anderen Fällen verwenden Kunden diese Parameter irrtümlich und das Ergebnis ist ein unnötiger Leistungsschaden.

Viele Kunden entfernen diese Mount-Optionen vorübergehend während der Installation oder dem Patching der Binärdateien der Anwendung. Diese Entfernung kann sicher durchgeführt werden, wenn der Benutzer überprüft, dass während der Installation oder des Patching-Prozesses keine anderen Prozesse aktiv das Zielverzeichnis verwenden.

NFS-Übertragungsgrößen

Standardmäßig beschränkt ONTAP die NFS-I/O-Größe auf 64K.

Zufälliger I/O mit den meisten Applikationen und Datenbanken verwendet eine viel kleinere Blockgröße, die weit unter dem 64K-Maximum liegt. Der I/O großer Blöcke wird in der Regel parallelisiert, sodass die 64K-Maximalgröße auch keine Einschränkung für die Erzielung der maximalen Bandbreite darstellt.

Es gibt einige Workloads, bei denen das 64K-Maximum eine Einschränkung darstellt. Insbesondere Vorgänge in einem einzigen Thread, wie Backup- oder Recovery-Vorgänge oder ein vollständiger Tabellenscan in einer Datenbank, laufen schneller und effizienter, wenn die Datenbank weniger, aber größere I/OS ausführen kann. Die optimale I/O-Handhabungsgröße für ONTAP beträgt 256 KB.

Die maximale Übertragungsgröße für eine bestimmte ONTAP SVM kann wie folgt geändert werden:

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

> Verringern Sie niemals die maximal zulässige Übertragungsgröße auf ONTAP unter den Wert rsize/wsize der aktuell gemounteten NFS-Dateisysteme. Dies kann bei einigen Betriebssystemen zu Hängebleiben oder sogar Datenbeschädigungen führen. Wenn beispielsweise NFS-Clients derzeit auf 65536 rsize/wsize gesetzt sind, dann könnte die maximale Übertragungsgröße für ONTAP ohne Auswirkung auf die Clients selbst begrenzt werden, zwischen 65536 und 1048576 angepasst werden. Wenn Sie die maximale Übertragungsgröße unter 65536 verringern, können die Verfügbarkeit oder die Daten beeinträchtigt werden.

NV-FEHLER

NVFAIL ist eine Funktion von ONTAP, die in katastrophalen Failover-Szenarien die Integrität sicherstellt.

Datenbanken sind bei Storage Failover-Ereignissen anfällig für Beschädigungen, da große interne Caches verfügbar sind. Wenn ein katastrophales Ereignis das Erzwingen eines ONTAP-Failovers oder das Erzwingen einer MetroCluster-Umschaltung erfordert, kann das Ergebnis, unabhängig vom Zustand der Gesamtkonfiguration, effektiv verworfen werden. Der Inhalt des Storage-Arrays springt zurück in die Zeit, und der Status des Datenbank-Cache entspricht nicht mehr dem Status der Daten auf der Festplatte. Diese Inkonsistenz führt zu Datenbeschädigung.

Caching kann auf Applikations- oder Serverebene erfolgen. Beispielsweise werden in einer Oracle RAC-Konfiguration (Real Application Cluster) mit Servern, die sowohl auf einem primären Standort als auch an einem Remote-Standort aktiv sind, Daten innerhalb des Oracle SGA zwischengespeichert. Bei einem erzwungenen Switchover-Vorgang, der zu einem Datenverlust führte, würde die Datenbank beschädigt werden, da die im SGA gespeicherten Blöcke möglicherweise nicht mit den Blöcken auf der Festplatte übereinstimmen.

Eine weniger offensichtliche Verwendung von Caching erfolgt auf der Ebene des Betriebssystems. Blöcke aus einem gemounteten NFS-Filesystem können im OS zwischengespeichert werden. Alternativ kann ein geclustertes Filesystem, das auf LUNs am primären Standort basiert, auf Servern am Remote-Standort gemountet werden, und wieder einmal konnten Daten zwischengespeichert werden. Ein Ausfall von NVRAM oder eine erzwungene Übernahme oder ein erzwungenes Switchover kann in diesen Situationen zu einer Beschädigung des File-Systems führen.

ONTAP schützt Datenbanken und Betriebssysteme vor diesem Szenario mit NVFAIL und den zugehörigen Einstellungen.

ASM Reclamation Utility (ASMRU)

Bei aktivierter Inline-Komprimierung entfernt ONTAP effizient Blöcke, die auf Dateien oder LUNs geschrieben werden, auf Null gesetzt. Dienstprogramme wie das Oracle ASM Reclamation Utility (ASRU) schreiben Nullen in ungenutzte ASM-Extents.

Auf diese Weise können DBAs nach dem Löschen von Daten Speicherplatz im Storage-Array zurückgewinnen. ONTAP fängt die Nullen ab und hebt den Speicherplatz von der LUN ab. Die Rückgewinnung erfolgt äußerst schnell, da innerhalb des Storage-Systems keine Daten geschrieben werden.

Aus der Datenbankperspektive enthält die ASM-Datenträgergruppe Nullen, und das Lesen dieser Bereiche der LUNs würde zu einem Strom von Nullen führen, ONTAP speichert die Nullen jedoch nicht auf Laufwerken. Stattdessen werden einfache Metadatenänderungen vorgenommen, die intern die Bereiche, in denen der Wert auf Null gesetzt wurde, als leer von Daten markieren.

Aus ähnlichen Gründen sind Performance-Tests mit gelöschten Daten nicht gültig, da Blöcke mit Nullen tatsächlich nicht als Schreibvorgänge innerhalb des Storage-Arrays verarbeitet werden.

Stellen Sie bei der Verwendung von ASRU sicher, dass alle von Oracle empfohlenen Patches installiert sind.

Storage-Konfiguration auf ASA r2-Systemen

FC SAN

LUN-Ausrichtung

LUN-Ausrichtung bezieht sich auf die I/O-Optimierung in Bezug auf das zugrunde liegende Filesystem-Layout.

ASA r2-Systeme nutzen die gleiche ONTAP Architektur wie AFF/ FAS , jedoch mit einem vereinfachten Konfigurationsmodell. ASA r2-Systeme verwenden Storage Availability Zones (SAZ) anstelle von Aggregaten, aber die Ausrichtungsprinzipien bleiben gleich, da ONTAP das Blocklayout plattformübergreifend konsistent verwaltet. Beachten Sie jedoch folgende ASA-spezifische Punkte:

- ASA r2-Systeme bieten aktiv-aktiv symmetrische Pfade für alle LUNs, wodurch Probleme mit Pfadasymmetrien während der Ausrichtung beseitigt werden.
- Speichereinheiten (LUNs) werden standardmäßig Thin-Provisioning-fähig gemacht; die Ausrichtung ändert nichts an diesem Verhalten.
- Snapshot-Reservierung und automatische Snapshot-Löschung können bei der LUN-Erstellung konfiguriert werden (ONTAP 9.18.1 und höher).

Auf einem ONTAP-System wird der Storage in 4-KB-Einheiten organisiert. Ein Datenbank- oder Filesystem-8-KB-Block sollte exakt zwei 4-KB-Blöcken zugeordnet werden. Wenn ein Fehler in der LUN-Konfiguration die Ausrichtung um 1 KB in beide Richtungen verschiebt, wäre jeder 8-KB-Block auf drei verschiedenen 4-KB-Storage-Blöcken vorhanden anstatt auf zwei. Diese Anordnung würde zu einer erhöhten Latenz führen und dazu führen, dass zusätzliche I/O-Vorgänge innerhalb des Speichersystems ausgeführt werden.

Die Ausrichtung wirkt sich auch auf LVM-Architekturen aus. Wenn ein physisches Volume innerhalb einer logischen Volume-Gruppe auf dem gesamten Laufwerk definiert wird (es werden keine Partitionen erstellt), wird der erste 4-KB-Block auf der LUN auf den ersten 4-KB-Block im Storage-System ausgerichtet. Dies ist eine korrekte Ausrichtung. Probleme ergeben sich bei Partitionen, da sie den Startort verschieben, an dem das Betriebssystem die LUN verwendet. Solange der Offset in ganzen 4-KB-Einheiten verschoben wird, ist die LUN ausgerichtet.

Erstellen Sie in Linux-Umgebungen logische Volume-Gruppen auf dem gesamten Laufwerkgerät. Wenn eine Partition erforderlich ist, überprüfen Sie die Ausrichtung, indem Sie ausführen `fdisk -u` und überprüfen, ob der Anfang jeder Partition ein Vielfaches von acht ist. Dies bedeutet, dass die Partition bei einem Vielfachen von acht 512-Byte-Sektoren beginnt, was 4 KB ist.

Siehe auch die Diskussion über die Ausrichtung der Kompressionsblöcke im Abschnitt "Effizienz". Jedes Layout, das an 8-KB-Komprimierungsblockgrenzen ausgerichtet ist, ist auch an 4-KB-Grenzen ausgerichtet.

Warnungen wegen Falschausrichtung

Die Datenbank-Wiederherstellungs-/Transaktionsprotokollierung erzeugt normalerweise nicht ausgerichtete I/O-Vorgänge, die irreführende Warnungen zu falsch ausgerichteten LUNs auf ONTAP verursachen können.

Die Protokollierung führt einen sequenziellen Schreibvorgang der Protokolldatei mit unterschiedlich großen Schreibvorgängen durch. Ein Protokollschreibvorgang, der sich nicht an 4-KB-Grenzen ausrichtet, verursacht normalerweise keine Performance-Probleme, da der nächste Protokollschreibvorgang den Block abgeschlossen hat. Das Ergebnis: ONTAP ist in der Lage, fast alle Schreibvorgänge als komplette 4-KB-Blöcke zu verarbeiten, obwohl die Daten in einigen 4-KB-Blöcken in zwei separaten Operationen geschrieben wurden.

Überprüfen Sie die Ausrichtung mithilfe von Hilfsprogrammen wie z. B. `sio` oder `dd` das E/A mit einer definierten Blockgröße erzeugen kann. Die I/O-Ausrichtungsstatistiken des Speichersystems können mit folgendem angezeigt werden: `stats` Befehl. Sehen "Überprüfung der WAFL-Ausrichtung" für weitere Informationen.

Die Ausrichtung in Solaris-Umgebungen ist komplizierter. Siehe "ONTAP SAN-Host-Konfiguration" Finden Sie weitere Informationen.

Achten Sie in Solaris x86-Umgebungen besonders auf die richtige Ausrichtung, da die meisten Konfigurationen mehrere Ebenen von Partitionen haben. Solaris x86-Partitionsschichten befinden sich in der Regel oben auf einer Standard-Master-Bootdatensammelpartitionstabelle.

Weitere bewährte Vorgehensweisen:

- Überprüfen Sie die Firmware- und Betriebssystemeinstellungen des HBA anhand des NetApp Interoperability Matrix Tool (IMT).
- Verwenden Sie die sanlun-Dienstprogramme, um die Integrität und Ausrichtung des Pfades zu überprüfen.
- Stellen Sie bei Oracle ASM und LVM sicher, dass die Konfigurationsdateien (/etc/lvm/lvm.conf, /etc/sysconfig/oracleasm) korrekt eingestellt sind, um Ausrichtungsprobleme zu vermeiden.

LUN-Dimensionierung und LUN-Anzahl

Die Auswahl der optimalen LUN-Größe und der Anzahl der zu verwendenden LUNs ist für optimale Performance und einfaches Management der Oracle-Datenbanken von entscheidender Bedeutung.

Eine LUN ist ein virtualisiertes Objekt auf ONTAP , das sich über alle Laufwerke in der hostenden Storage Availability Zone (SAZ) auf ASA r2-Systemen erstreckt. Daher wird die Leistungsfähigkeit der LUN nicht durch ihre Größe beeinträchtigt, da die LUN unabhängig von der gewählten Größe das volle Leistungspotenzial der SAZ ausschöpft.

Aus praktischen Gründen möchten Kunden möglicherweise eine LUN einer bestimmten Größe verwenden. Wenn beispielsweise eine Datenbank auf einer LVM oder einer Oracle ASM-Datenträgergruppe erstellt wird, die aus zwei LUNs mit jeweils 1 TB besteht, muss diese Datenträgergruppe in Schritten von 1 TB erweitert werden. Es könnte besser sein, die Datenträgergruppe aus acht LUNs mit jeweils 500 GB zu erstellen, damit die Datenträgergruppe in kleineren Schritten erhöht werden kann.

Die Praxis, eine universelle Standard-LUN-Größe zu etablieren, wird davon abgeraten, da dies die Managebarkeit erschweren kann. Beispielsweise funktioniert eine standardmäßige LUN-Größe von 100 GB gut, wenn eine Datenbank oder ein Datastore im Bereich von 1 TB bis 2 TB liegt, jedoch erfordert eine Datenbank oder ein Datenspeicher mit einer Größe von 20 TB 200 LUNs. Das bedeutet, dass der Server-Neustart länger dauert, mehr Objekte in den verschiedenen Benutzeroberflächen zu verwalten sind und Produkte wie SnapCenter eine Erkennung für viele Objekte durchführen müssen. Derartige Probleme werden durch die Verwendung von weniger und größeren LUNs vermieden.

- ASA r2-Überlegungen:*
- Die maximale LUN-Größe für ASA r2 beträgt 128 TB, was den Einsatz von weniger, aber größeren LUNs ohne Leistungseinbußen ermöglicht.
- ASA r2 verwendet Storage Availability Zones (SAZ) anstelle von Aggregaten, dies ändert jedoch nichts an der Logik zur LUN-Dimensionierung für Oracle-Workloads.
- Thin Provisioning ist standardmäßig aktiviert; die Größenänderung von LUNs erfolgt unterbrechungsfrei und erfordert keine Offline-Schaltung.

LUN-Anzahl

Anders als die LUN-Größe wirkt sich die Anzahl der LUNs auf die Performance aus. Die Applikations-Performance hängt häufig von der Fähigkeit ab, parallelen I/O über die SCSI-Schicht auszuführen. Dadurch bieten zwei LUNs eine bessere Performance als eine einzelne LUN. Die Verwendung einer LVM wie Veritas VxVM, Linux LVM2 oder Oracle ASM ist die einfachste Methode, um die Parallelität zu erhöhen.

Bei ASA r2 bleiben die Prinzipien für die LUN-Anzahl die gleichen wie bei AFF/ FAS , da ONTAP parallele E/A plattformübergreifend ähnlich handhabt. Die SAN-only-Architektur und die aktiv-aktiven symmetrischen Pfade der ASA r2 gewährleisten jedoch eine konsistente Leistung über alle LUNs hinweg.

NetApp Kunden konnten im Allgemeinen nur einen minimalen Nutzen aus der Erhöhung der Anzahl von LUNs über sechzehn hinaus verzeichnen, obwohl sich bei den Tests mit 100 % SSD-Umgebungen mit sehr hoher zufälliger I/O-Last weitere Verbesserungen auf bis zu 64 LUNs gezeigt haben.

NetApp empfiehlt Folgendes:

Im Allgemeinen reichen vier bis sechzehn LUNs aus, um die E/A-Anforderungen einer beliebigen Oracle-Datenbank-Workload zu erfüllen. Weniger als vier LUNs könnten aufgrund von Einschränkungen bei den Host-SCSI-Implementierungen zu Leistungseinschränkungen führen. Eine Erhöhung auf mehr als sechzehn LUNs führt nur in Extremfällen (z. B. bei sehr hohen zufälligen E/A-SSD-Workloads) zu einer Leistungsverbesserung.

LUN-Platzierung

Die optimale Platzierung von Datenbank-LUNs innerhalb von ASA r2-Systemen hängt in erster Linie davon ab, wie die verschiedenen ONTAP Funktionen genutzt werden.

In ASA r2-Systemen werden Speichereinheiten (LUNs oder NVMe-Namespaces) aus einer vereinfachten Speicherschicht namens Storage Availability Zones (SAZs) erstellt, die als gemeinsame Speicherpools für ein HA-Paar fungieren.

Pro HA-Paar gibt es typischerweise nur eine Storage Availability Zone (SAZ).

Speicherverfügbarkeitszonen (SAZ)

In ASA r2-Systemen sind Volumes weiterhin vorhanden, werden aber automatisch erstellt, wenn Speichereinheiten angelegt werden. Speichereinheiten (LUNs oder NVMe-Namespaces) werden direkt in den automatisch erstellten Volumes in Storage Availability Zones (SAZs) bereitgestellt. Dieses Design macht eine manuelle Volumenverwaltung überflüssig und ermöglicht eine direktere und effizientere Bereitstellung für Block-Workloads wie Oracle-Datenbanken.

SAZs und Lagereinheiten

Zusammengehörige Speichereinheiten (LUNs oder NVMe-Namensräume) befinden sich normalerweise innerhalb einer einzigen Storage Availability Zone (SAZ). Eine Datenbank, die beispielsweise 10 Speichereinheiten (LUNs) benötigt, würde typischerweise alle 10 Einheiten aus Gründen der Einfachheit und Leistungsfähigkeit in derselben SAZ platzieren.

- Die Verwendung eines 1:1-Verhältnisses von Speichereinheiten zu Datenträgern, d. h. eine Speichereinheit (LUN) pro Datenträger, ist das Standardverhalten von ASA r2.
- Falls im ASA r2-System mehr als ein HA-Paar vorhanden ist, können die Speichereinheiten (LUNs) für eine bestimmte Datenbank auf mehrere SAZs verteilt werden, um die Controller-Auslastung und -Leistung zu optimieren.

Im Kontext von FC SAN bezieht sich die Speichereinheit hier auf eine LUN.

Konsistenzgruppen (CGs), LUNs und Snapshots

In ASA r2 werden Snapshot-Richtlinien und -Zeitpläne auf der Ebene der Konsistenzgruppe angewendet. Dabei handelt es sich um ein logisches Konstrukt, das mehrere LUNs oder NVMe-Namespaces zum Zweck des koordinierten Datenschutzes gruppiert. Ein Datensatz, der aus 10 LUNs besteht, benötigt nur eine einzige Snapshot-Richtlinie, wenn diese LUNs Teil derselben Konsistenzgruppe sind.

Konsistenzgruppen gewährleisten atomare Snapshot-Operationen über alle einbezogenen LUNs hinweg. Beispielsweise kann eine Datenbank, die sich auf 10 LUNs befindet, oder eine VMware-basierte Anwendungsumgebung, die aus 10 verschiedenen Betriebssystemen besteht, als ein einziges, konsistentes

Objekt geschützt werden, wenn die zugrunde liegenden LUNs in derselben Konsistenzgruppe gruppiert sind. Werden sie in verschiedenen Konsistenzgruppen platziert, können Snapshots perfekt synchronisiert sein oder auch dann nicht, wenn sie gleichzeitig geplant werden.

In einigen Fällen muss ein zusammengehöriger Satz von LUNs aufgrund von Wiederherstellungsanforderungen möglicherweise in zwei verschiedene Konsistenzgruppen aufgeteilt werden. Eine Datenbank könnte beispielsweise vier LUNs für Datendateien und zwei LUNs für Protokolle haben. In diesem Fall wäre eine Datendatei-Konsistenzgruppe mit 4 LUNs und eine Protokoll-Konsistenzgruppe mit 2 LUNs möglicherweise die beste Option. Der Grund dafür ist die unabhängige Wiederherstellbarkeit: Die Datendatei-Konsistenzgruppe könnte selektiv auf einen früheren Zustand zurückgesetzt werden, was bedeutet, dass alle vier LUNs auf den Zustand des Snapshots zurückgesetzt würden, während die Protokoll-Konsistenzgruppe mit ihren kritischen Daten unberührt bliebe.

CGs, LUNs und SnapMirror

SnapMirror Richtlinien und -Operationen werden, wie Snapshot-Operationen, auf der Konsistenzgruppe und nicht auf der LUN ausgeführt.

Durch die Zusammenlegung verwandter LUNs in einer einzigen Konsistenzgruppe können Sie eine einzige SnapMirror Beziehung erstellen und alle enthaltenen Daten mit einem einzigen Update aktualisieren. Wie bei Snapshots handelt es sich auch bei der Aktualisierung um eine atomare Operation. Das SnapMirror Zielsystem würde garantiert eine exakte, zeitpunktbezogene Replik der Quell-LUNs enthalten. Wenn die LUNs über mehrere Konsistenzgruppen verteilt sind, können die Replikate untereinander konsistent sein oder auch nicht.

Die SnapMirror Replikation auf ASA r2-Systemen weist folgende Einschränkungen auf:

- Die synchrone Replikation von SnapMirror wird nicht unterstützt.
- SnapMirror Active Sync wird nur zwischen zwei ASA r2-Systemen unterstützt.
- SnapMirror unterstützt die asynchrone Replikation nur zwischen zwei ASA r2-Systemen.
- Die asynchrone Replikation von SnapMirror wird zwischen einem ASA r2-System und einem ASA, AFF oder FAS System oder der Cloud nicht unterstützt.

Erfahren Sie mehr über "SnapMirror Replikationsrichtlinien werden auf ASA r2-Systemen unterstützt"Die

CGs, LUNs und QoS

QoS kann zwar selektiv auf einzelne LUNs angewendet werden, in der Regel ist es jedoch einfacher, es auf der Ebene der Konsistenzgruppe festzulegen. Beispielsweise könnten alle von den Gästen auf einem bestimmten ESX-Server verwendeten LUNs in einer einzigen Konsistenzgruppe zusammengefasst werden, und anschließend könnte eine adaptive QoS-Richtlinie von ONTAP angewendet werden. Das Ergebnis ist ein sich selbst skalierender IOPS-pro-TiB-Grenzwert, der für alle LUNs gilt.

Wenn beispielsweise eine Datenbank 100.000 IOPS benötigt und 10 LUNs belegt, wäre es einfacher, ein einzelnes Limit von 100.000 IOPS für eine einzelne Konsistenzgruppe festzulegen, als 10 einzelne Limits von jeweils 10.000 IOPS, eines für jede LUN.

Mehrere CG-Layouts

Es gibt Fälle, in denen die Verteilung von LUNs auf mehrere Konsistenzgruppen von Vorteil sein kann. Der Hauptgrund ist die Controller-Striping-Funktion. Beispielsweise könnte ein HA ASA r2-Speichersystem eine einzelne Oracle-Datenbank hosten, bei der das volle Verarbeitungs- und Caching-Potenzial jedes Controllers benötigt wird. In diesem Fall wäre ein typisches Design, die Hälfte der LUNs in einer einzigen

Konsistenzgruppe auf Controller 1 und die andere Hälfte der LUNs in einer einzigen Konsistenzgruppe auf Controller 2 zu platzieren.

In ähnlicher Weise kann in Umgebungen mit vielen Datenbanken durch die Verteilung von LUNs auf mehrere Konsistenzgruppen eine ausgewogene Controller-Auslastung sichergestellt werden. Ein HA-System, das beispielsweise 100 Datenbanken mit jeweils 10 LUNs hostet, könnte pro Datenbank 5 LUNs einer Konsistenzgruppe auf Controller 1 und 5 LUNs einer Konsistenzgruppe auf Controller 2 zuweisen. Dies gewährleistet eine symmetrische Auslastung bei der Bereitstellung zusätzlicher Datenbanken.

Keines dieser Beispiele beinhaltet jedoch ein LUN-zu-Konsistenzgruppen-Verhältnis von 1:1. Ziel bleibt es, die Verwaltbarkeit zu optimieren, indem zusammengehörige LUNs logisch in Konsistenzgruppen zusammengefasst werden.

Ein Beispiel, bei dem ein Verhältnis von 1:1 zwischen LUN und Konsistenzgruppe sinnvoll ist, sind containerisierte Workloads. Hierbei repräsentiert jede LUN möglicherweise einen einzelnen Workload, der separate Snapshot- und Replikationsrichtlinien erfordert und daher individuell verwaltet werden muss. In solchen Fällen kann ein Verhältnis von 1:1 optimal sein.

LUN-Größe und LVM-Größe

Wenn ein SAN-basiertes Dateisystem oder eine Oracle ASM-Disk-Gruppe auf ASA r2 ihre Kapazitätsgrenze erreicht, gibt es zwei Möglichkeiten, den verfügbaren Speicherplatz zu erhöhen:

- Vergrößern Sie die Größe der vorhandenen LUNs (Speichereinheiten).
- Fügen Sie einer vorhandenen ASM-Datenträgergruppe oder LVM-Volumegruppe eine neue LUN hinzu und vergrößern Sie das enthaltene logische Volume.

Obwohl die Größenänderung von LUNs auf ASA r2 unterstützt wird, ist es im Allgemeinen besser, einen Logical Volume Manager (LVM) wie Oracle ASM zu verwenden. Einer der Hauptgründe für die Existenz von LVMs ist die Vermeidung der Notwendigkeit häufiger LUN-Größenänderungen. Bei einem LVM werden mehrere LUNs zu einem virtuellen Speicherpool zusammengefasst. Aus diesem Pool erstellte logische Volumes können problemlos in ihrer Größe angepasst werden, ohne die zugrunde liegende Speicherkonfiguration zu beeinträchtigen.

Weitere Vorteile der Verwendung von LVM oder ASM sind:

- Leistungsoptimierung: Verteilt die E/A auf mehrere LUNs und reduziert so Hotspots.
- Flexibilität: Neue LUNs hinzufügen, ohne bestehende Workloads zu unterbrechen.
- Transparente Migration: ASM oder LVM können Extents ohne Host-Ausfallzeit auf neue LUNs verschieben, um Lastausgleich oder Tiering zu ermöglichen.

Wichtige Überlegungen gemäß ASA r2:

- Die Größenänderung der LUNs erfolgt auf Ebene der Speichereinheit innerhalb einer Storage VM (SVM) unter Verwendung der Kapazität der Storage Availability Zone (SAZ).
- Für Oracle ist es Best Practice, LUNs zu ASM-Disk-Gruppen hinzuzufügen, anstatt bestehende LUNs zu vergrößern oder zu verkleinern, um Striping und Parallelität zu erhalten.

LVM-Striping

LVM-Striping bezieht sich auf die Verteilung von Daten über mehrere LUNs. So lässt sich die Performance vieler Datenbanken deutlich steigern.

Vor der Ära der Flash-Laufwerke wurde Striping verwendet, um die Performance-Einschränkungen rotierender Laufwerke zu überwinden. Beispiel: Wenn ein Betriebssystem einen Lesevorgang von 1 MB ausführen muss, würde das Lesen dieser 1 MB Daten von einem einzigen Laufwerk viel Festplattenkopf erfordern, der sucht und liest, da die 1 MB langsam übertragen wird. Wenn diese 1 MB Daten über 8 LUNs verteilt wurden, kann das Betriebssystem acht 128K-Lesevorgänge parallel ausführen und die für die 1-MB-Übertragung erforderliche Zeit verringern.

Das Striping mit rotierenden Laufwerken war schwieriger, da das I/O-Muster im Voraus bekannt sein musste. Wenn die Streifenbildung nicht korrekt auf die tatsächlichen I/O-Muster abgestimmt ist, kann dies die Leistung beeinträchtigen. Bei Oracle-Datenbanken, insbesondere bei All-Flash-Speicherkonfigurationen, ist Striping wesentlich einfacher zu konfigurieren und verbessert die Leistung nachweislich dramatisch.

Logische Volume-Manager wie Oracle ASM Stripe sind standardmäßig aktiviert, aber native OS LVM nicht. Einige von ihnen verbinden mehrere LUNs als verkettete Geräte. Dies führt zu Datendateien, die auf einem und nur einem LUN-Gerät vorhanden sind. Dies verursacht Hotspots. Andere LVM-Implementierungen sind standardmäßig auf verteilte Extents eingestellt. Das ist ähnlich wie Striping, aber es ist gröber. Die LUNs in der Volume-Gruppe werden in große Teile geteilt, die als Extents bezeichnet werden und in der Regel in vielen Megabyte gemessen werden. Die logischen Volumes werden dann über diese Extents verteilt. Das Ergebnis ist ein zufälliger I/O-Vorgang für eine Datei, der auf LUNs verteilt werden sollte. Sequenzielle I/O-Vorgänge sind jedoch nicht so effizient wie möglich.

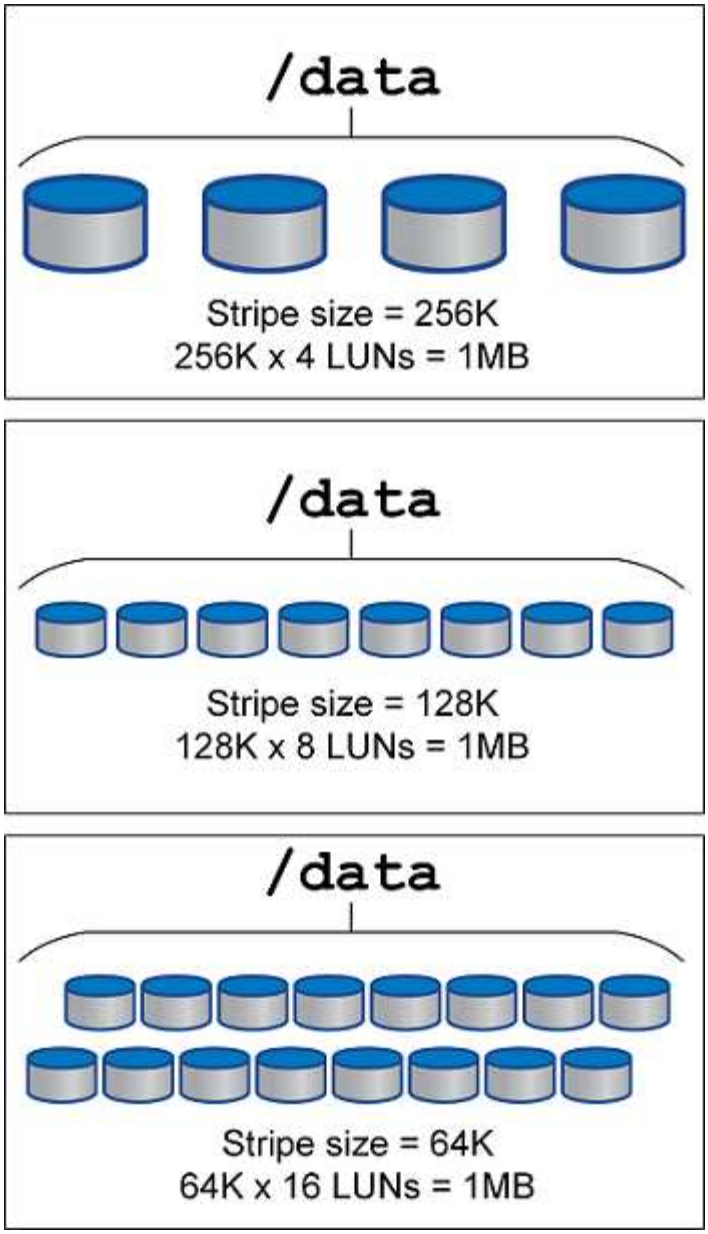
Die Performance-intensiven Applikations-I/O-Vorgänge erfolgen fast immer entweder (a) in Einheiten der grundlegenden Blockgröße oder (b) in Megabyte.

Das primäre Ziel einer Striped-Konfiguration ist es, sicherzustellen, dass Single-File I/O als eine Einheit ausgeführt werden kann. Multiblock-I/O, die eine Größe von 1 MB haben sollte, kann gleichmäßig über alle LUNs im Striped Volume hinweg parallelisiert werden. Das bedeutet, dass die Stripe-Größe nicht kleiner als die Blockgröße der Datenbank sein darf und die Stripe-Größe multipliziert mit der Anzahl der LUNs 1 MB betragen sollte.

Bewährte Vorgehensweise für LVM-Striping mit Oracle-Datenbanken:

- Stripe-Größe ≥ Datenbank-Blockgröße.
- Stripe-Größe * Anzahl der LUNs ≈ 1 MB für optimale Parallelität.
- Um den Durchsatz zu maximieren und Hotspots zu vermeiden, sollten mehrere LUNs pro ASM-Disk-Gruppe verwendet werden.

Die folgende Abbildung zeigt drei mögliche Optionen für die Stripe-Größe und Breitenabstimmung. Die Anzahl der LUNs wird ausgewählt, um die oben beschriebenen Performance-Anforderungen zu erfüllen. In allen Fällen beträgt die Gesamtzahl der Daten innerhalb eines einzigen Stripes jedoch 1 MB.



NV-FEHLER

NVFAIL ist eine ONTAP Funktion, die die Datenintegrität bei katastrophalen Ausfallszenarien gewährleistet.

Diese Funktionalität ist auch auf ASA r2-Systemen anwendbar, obwohl ASA r2 eine vereinfachte SAN-Architektur verwendet (SAZs und Speichereinheiten anstelle von Volumes).

Datenbanken sind bei Speicherausfallereignissen anfällig für Datenbeschädigung, da sie große interne Caches verwalten. Wenn ein katastrophales Ereignis ein erzwungenes ONTAP Failover erfordert, unabhängig vom Zustand der Gesamtkonfiguration, kann dies dazu führen, dass zuvor vorgenommene Änderungen effektiv verworfen werden. Der Inhalt des Speicherarrays springt zeitlich zurück, und der Zustand des Datenbankcaches spiegelt nicht mehr den Zustand der Daten auf der Festplatte wider. Diese Inkonsistenz führt zu Datenbeschädigung.

Caching kann auf Anwendungs- oder Serverebene erfolgen. Beispielsweise speichert eine Oracle Real Application Cluster (RAC)-Konfiguration mit Servern, die sowohl an einem primären als auch an einem

Remote-Standort aktiv sind, Daten im Oracle SGA zwischen. Ein erzwungener Failover-Vorgang, der zu Datenverlusten führen würde, würde die Datenbank der Gefahr einer Beschädigung aussetzen, da die im SGA gespeicherten Blöcke möglicherweise nicht mit den Blöcken auf der Festplatte übereinstimmen.

Eine weniger offensichtliche Anwendung des Caching findet sich auf der Ebene des Betriebssystem-Dateisystems. Ein auf LUNs basierendes Cluster-Dateisystem, das sich am primären Standort befindet, könnte auf Servern am entfernten Standort eingebunden werden, und auch hier könnten Daten zwischengespeichert werden. Ein Ausfall des NVRAM oder eine erzwungene Übernahme in diesen Situationen könnten zu einer Beschädigung des Dateisystems führen.

ONTAP schützt Datenbanken und Betriebssysteme vor diesem Szenario mithilfe von NVFAIL und den zugehörigen Einstellungen, die dem Host signalisieren, zwischengespeicherte Daten zu invalidieren und die betroffenen Dateisysteme nach einem Failover neu einzubinden. Dieser Mechanismus gilt für ASA r2 LUNs und Namespaces genauso wie für AFF/ FAS.

Wichtige Überlegungen gemäß ASA r2:

- NVFAIL arbeitet auf LUN-Ebene (Speichereinheit), nicht auf SAZ-Ebene.
- Bei Oracle-Datenbanken sollte NVFAIL auf allen LUNs aktiviert werden, die kritische Komponenten (Datendateien, Redo-Logs, Kontrolldateien) hosten.
- MetroCluster wird auf ASA r2 nicht unterstützt, daher ist NVFAIL hauptsächlich für lokale HA-Failover-Szenarien relevant.
- NFS wird auf ASA r2 nicht unterstützt, daher gelten die NVFAIL-Überlegungen nur für SAN-basierte Workloads (FC/iSCSI/NVMe).

ASM-Rückgewinnungsdienstprogramm (ASRU)

ONTAP auf ASA r2 entfernt effizient Nullblöcke, die auf eine LUN (Speichereinheit) geschrieben wurden, wenn die Inline-Komprimierung aktiviert ist. Dienstprogramme wie das Oracle ASM Reclamation Utility (ASRU) funktionieren, indem sie ungenutzte ASM-Extents mit Nullen überschreiben.

Dies ermöglicht es Datenbankadministratoren, Speicherplatz auf dem Speichersystem freizugeben, nachdem Daten gelöscht wurden. ONTAP fängt die Nullen ab und gibt den Speicherplatz der LUN frei. Der Rückgewinnungsprozess ist extrem schnell, da keine Daten im Speichersystem geschrieben werden.

Aus der Datenbankperspektive enthält die ASM-Datenträgergruppe Nullen, und das Lesen dieser Bereiche der LUNs würde zu einem Strom von Nullen führen, ONTAP speichert die Nullen jedoch nicht auf Laufwerken. Stattdessen werden einfache Metadatenänderungen vorgenommen, die intern die Bereiche, in denen der Wert auf Null gesetzt wurde, als leer von Daten markieren.

Aus ähnlichen Gründen sind Performance-Tests mit gelöschten Daten nicht gültig, da Blöcke mit Nullen tatsächlich nicht als Schreibvorgänge innerhalb des Storage-Arrays verarbeitet werden.

Wichtige ASRU-Überlegungen mit ASA r2 ONTAP:

- Funktioniert bei SAN-Workloads genauso wie AFF/ FAS , da ASA r2 nur blockbasiert arbeitet.
- Gilt für LUNs und NVMe-Namespaces, die innerhalb von SAZs bereitgestellt werden.
- Es existieren keine FlexVol Volumes, aber das Verhalten bei der Nullblock-Rückgewinnung ist identisch.

Stellen Sie bei der Verwendung von ASRU sicher, dass alle von Oracle empfohlenen Patches installiert sind.

Einheitliche

Die Virtualisierung von Datenbanken mit VMware, Oracle OLVM oder KVM wird für NetApp-Kunden, die sich für Virtualisierung selbst für ihre geschäftskritischsten Datenbanken entschieden haben, immer häufiger eingesetzt.

Instandhaltung

Es gibt viele Missverständnisse bezüglich der Oracle Support-Richtlinien für Virtualisierung, insbesondere für VMware-Produkte. Es ist nicht ungewöhnlich, zu hören, dass Oracle die Virtualisierung nicht unterstützt. Diese Vorstellung ist falsch und führt zu verpassten Möglichkeiten, von der Virtualisierung zu profitieren. Oracle Doc ID 249212.1 behandelt die tatsächlichen Anforderungen und wird von Kunden selten als ein Problem betrachtet.

Wenn ein Problem auf einem virtualisierten Server auftritt und das Problem dem Oracle Support bisher unbekannt war, wird der Kunde möglicherweise gebeten, das Problem auf physischer Hardware zu reproduzieren. Ein Oracle Kunde, der eine innovative Version eines Produkts ausführt, möchte möglicherweise wegen möglicher Supportprobleme keine Virtualisierung nutzen. Für Virtualisierungskunden, die allgemein verfügbare Oracle Produktversionen verwenden, war diese Situation jedoch keine echte Realität.

Storage-Präsentation

Kunden, die eine Virtualisierung ihrer Datenbanken in Erwägung ziehen, sollten ihre Storage-Entscheidungen basierend auf ihren Geschäftsanforderungen treffen. Dies ist zwar für alle IT-Entscheidungen generell eine echte Aussage, ist aber vor allem bei Datenbankprojekten von Bedeutung, da sich Größe und Umfang der Anforderungen erheblich unterscheiden.

Es gibt drei grundlegende Optionen für die Storage-Präsentation:

- Virtualisierte LUNs auf Hypervisor-Datastores
- Vom iSCSI-Initiator gemanagte iSCSI-LUNs auf der VM, nicht der Hypervisor
- NFS-Dateisysteme, die von der VM gemountet wurden (nicht aus einem NFS-basierten Datastore)
- Direkte Gerätezuordnungen. VMware RDMs werden von Kunden missbWege, aber physische Geräte werden immer noch oft ähnlich direkt mit KVM- und OLVM-Virtualisierung abgebildet.

Leistung

Die Methode zur Bereitstellung von Speicher für einen virtualisierten Gast hat in der Regel keine Auswirkungen auf die Leistung. Host-Betriebssysteme, virtualisierte Netzwerktreiber und Hypervisor-Datastore-Implementierungen sind allesamt hochoptimiert und können im Allgemeinen die gesamte verfügbare FC- oder IP-Netzwerkbandbreite zwischen dem Hypervisor und dem Storage-System nutzen, sofern grundlegende Best Practices befolgt werden. In einigen Fällen ist das Erlangen einer optimalen Performance unter Umständen mit einem Ansatz für die Storage-Präsentation im Vergleich zu einem anderen etwas einfacher, das Endergebnis sollte jedoch vergleichbar sein.

Managebarkeit

Der wichtigste Faktor bei der Entscheidung, wie Storage einem virtualisierten Gast zur Verfügung gestellt wird, ist die Fähigkeit, zu bestimmen. Es gibt keine richtige oder falsche Methode. Der beste Ansatz hängt von den Anforderungen, Fähigkeiten und Präferenzen der IT-Abteilung ab.

Zu den berücksichtigenden Faktoren gehören:

- **Transparenz.** Wenn eine VM ihre Dateisysteme verwaltet, ist es für einen Datenbankadministrator oder einen Systemadministrator einfacher, die Quelle der Dateisysteme für ihre Daten zu identifizieren. Auf die Dateisysteme und LUNs wird nicht anders zugegriffen als auf einem physischen Server.
- **Konsistenz.** Wenn eine VM Eigentümer ihrer Dateisysteme ist, wirkt sich die Verwendung oder Nichtnutzung einer Hypervisor-Schicht auf die Verwaltbarkeit aus. Die gleichen Verfahren für die Bereitstellung, das Monitoring, die Datensicherung usw. können über den gesamten Bestand hinweg eingesetzt werden, einschließlich sowohl virtualisierter als auch nicht virtualisierter Umgebungen.

  Andererseits könnte es in einem ansonsten zu 100 % virtualisierten Datacenter besser sein, Datastore-basierten Storage über den gesamten Platzbedarf hinweg zu nutzen – mit derselben Argumentation wie oben erwähnt – Konsistenz – die Fähigkeit, dieselben Verfahren für Bereitstellung, Sicherung, Überwachung und Datensicherung zu verwenden.

- **Stabilität und Fehlerbehebung.** Wenn eine VM über ihre Dateisysteme verfügt, sind die Bereitstellung guter, stabiler Performance und Fehlerbehebungsprobleme einfacher, da der gesamte Speicher-Stack auf der VM vorhanden ist. Die einzige Rolle des Hypervisors besteht darin, FC- oder IP-Frames zu transportieren. Wenn ein Datastore in einer Konfiguration enthalten ist, wird die Konfiguration durch die Einführung eines weiteren Satzes von Timeouts, Parametern, Protokolldateien und potenziellen Bugs erschwert.
- **Portabilität.** Wenn eine VM Eigentümer ihrer Dateisysteme ist, wird der Prozess des Verschiebens einer Oracle-Umgebung viel einfacher. Dateisysteme können problemlos zwischen virtualisierten und nicht virtualisierten Gastsystemen verschoben werden.
- **Vendor Lock-in.** Nachdem die Daten in einem Datastore platziert wurden, wird es schwierig, einen anderen Hypervisor zu verwenden oder die Daten aus der virtualisierten Umgebung zu nehmen.
- **Snapshot-Aktivierung.** herkömmliche Backup-Verfahren in einer virtualisierten Umgebung können wegen der relativ begrenzten Bandbreite zu einem Problem werden. Beispielsweise ist ein 10-GbE-Trunk mit vier Ports möglicherweise ausreichend, um den täglichen Leistungsbedarf vieler virtualisierter Datenbanken zu decken. Ein solcher Trunk wäre jedoch nicht ausreichend, um Backups mit RMAN oder anderen Backup-Produkten durchzuführen, die das Streaming einer vollständigen Kopie der Daten erfordern. So müssen in einer zunehmend konsolidierten virtualisierten Umgebung Backups über Storage Snapshots durchgeführt werden. Dadurch entfällt die Notwendigkeit, die Hypervisor-Konfiguration lediglich zu überbauen, um die Bandbreiten- und CPU-Anforderungen im Backup-Fenster zu unterstützen.

  Bei der Verwendung von Filesystemen, die sich im Besitz von Gästen befinden, ist es manchmal einfacher, Snapshot-basierte Backups und Restores zu nutzen, da die zu schützenden Storage-Objekte einfacher angepeißt werden können. Es gibt jedoch eine immer größere Anzahl an Datensicherungsprodukten für die Virtualisierung, die sich gut in Datenspeicher und Snapshots integrieren lassen. Die Backup-Strategie sollte vollständig berücksichtigt werden, bevor eine Entscheidung darüber getroffen wird, wie Speicher einem virtualisierten Host zur Verfügung gestellt werden kann.

Paravirtualisierte Treiber

Für eine optimale Leistung ist die Verwendung paravirtualisierter Netzwerktreiber von entscheidender Bedeutung. Wenn ein Datastore verwendet wird, ist ein paravirtualisierter SCSI-Treiber erforderlich. Ein

paravirtualisierter Gerätetreiber ermöglicht es einem Gast, sich tiefer in den Hypervisor zu integrieren, im Gegensatz zu einem emulierten Treiber, bei dem der Hypervisor mehr CPU-Zeit damit verbringt, das Verhalten physischer Hardware nachzuahmen.

RAM überschreiben

Das Überschreiben von RAM bedeutet, mehr virtualisierten RAM auf verschiedenen Hosts zu konfigurieren, als auf der physischen Hardware vorhanden ist. Dies kann zu unerwarteten Leistungsproblemen führen. Bei der Virtualisierung einer Datenbank dürfen die zugrunde liegenden Blöcke des Oracle SGA nicht durch den Hypervisor in den Storage getauscht werden. Dies führt zu äußerst instabilen Performance-Ergebnissen.

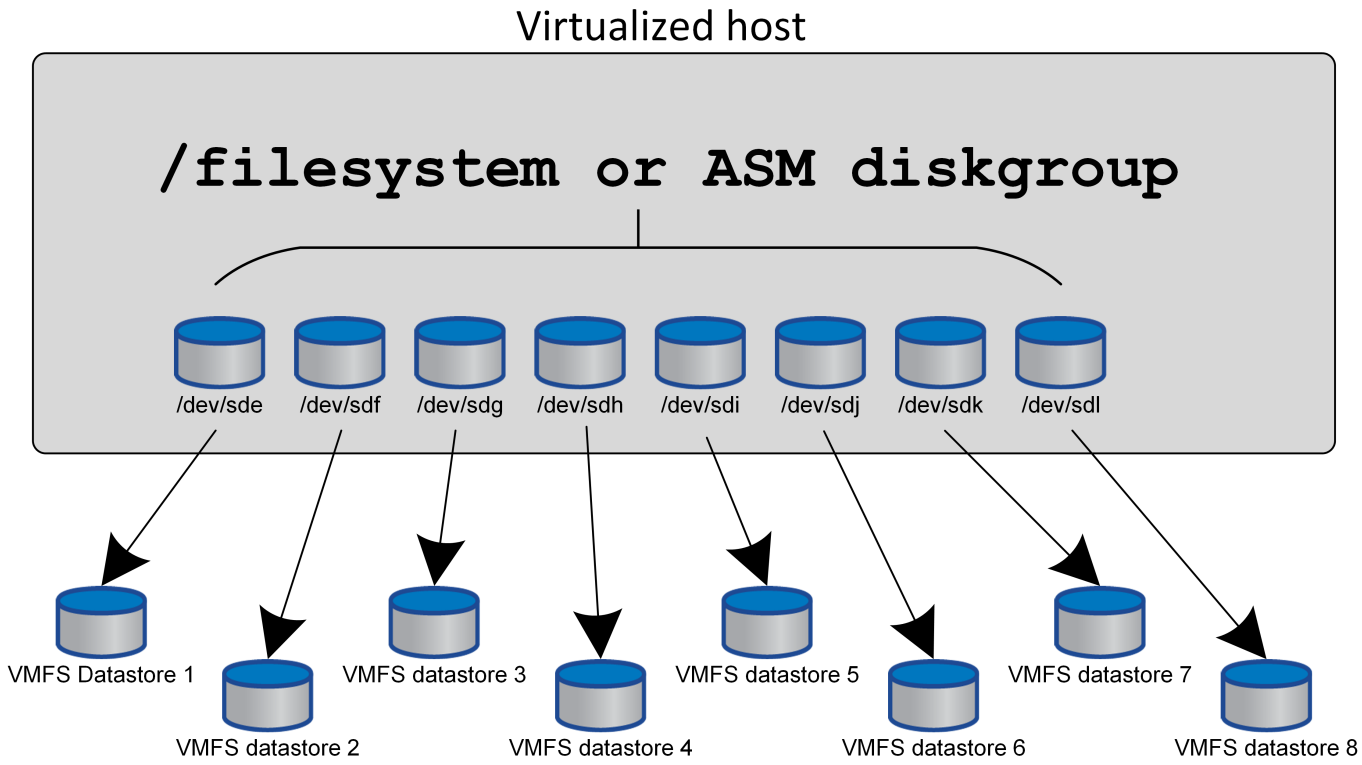
Datastore-Striping

Bei der Verwendung von Datenbanken mit Datastores muss ein entscheidender Faktor im Hinblick auf die Performance berücksichtigt werden: Striping.

Datastore-Technologien wie VMFS können mehrere LUNs umfassen, sind jedoch keine Striping-Geräte. Die LUNs werden verkettet. Das Endergebnis kann LUN-Hotspots sein. Beispielsweise könnte eine typische Oracle-Datenbank eine ASM-Festplattengruppe mit 8 LUNs enthalten. Alle 8 virtualisierten LUNs konnten auf einem VMFS Datenspeicher mit 8 LUNs bereitgestellt werden, es gibt jedoch keine Garantie, auf welchen LUNs sich die Daten befinden. Die resultierende Konfiguration könnte alle 8 virtualisierten LUNs sein, die eine einzelne LUN innerhalb des VMFS-Datastore belegen. Das führt zu einem Performance-Engpass.

Striping ist in der Regel erforderlich. Bei einigen Hypervisoren, einschließlich KVM, ist es möglich, wie beschrieben mit LVM Striping einen Datenspeicher zu erstellen "Hier". Bei VMware sieht die Architektur etwas anders aus. Jede virtualisierte LUN muss in einem anderen VMFS-Datenspeicher platziert werden.

Beispiel:



Der Haupttreiber dieses Ansatzes ist nicht ONTAP, sondern er liegt an der inhärenten Beschränkung der

Anzahl der Vorgänge, die eine einzelne VM oder Hypervisor-LUN parallel bedienen kann. Eine einzelne ONTAP-LUN kann im Allgemeinen deutlich mehr IOPS unterstützen, als ein Host anfordern kann. Das Performance-Limit für eine einzelne LUN ist fast universell ein Ergebnis des Host-Betriebssystems. Das Ergebnis: Die meisten Datenbanken benötigen zwischen 4 und 8 LUNs, um ihre Performance-Anforderungen zu erfüllen.

VMware Architekturen müssen ihre Architekturen sorgfältig planen, um sicherzustellen, dass sie keine Maxima für Datenspeicher und/oder LUN-Pfade aufweisen. Darüber hinaus ist keine Notwendigkeit für eine eindeutige Gruppe von VMFS-Datenspeichern für jede Datenbank erforderlich. Die primäre Anforderung besteht darin sicherzustellen, dass jeder Host über einen sauberen Satz von 4-8 I/O-Pfaden von den virtualisierten LUNs zu den Back-End-LUNs auf dem Speichersystem selbst verfügt. In seltenen Fällen können sogar noch mehr Daten für wirklich extreme Performance-Anforderungen von Vorteil sein, aber 4-8 LUNs sind im Allgemeinen für 95 % aller Datenbanken ausreichend. Ein einzelnes ONTAP Volume mit 8 LUNs kann bis zu 250,000 zufällige Oracle Block-IOPS mit einer typischen OS-/ONTAP-/Netzwerkkonfiguration unterstützen.

Tiering

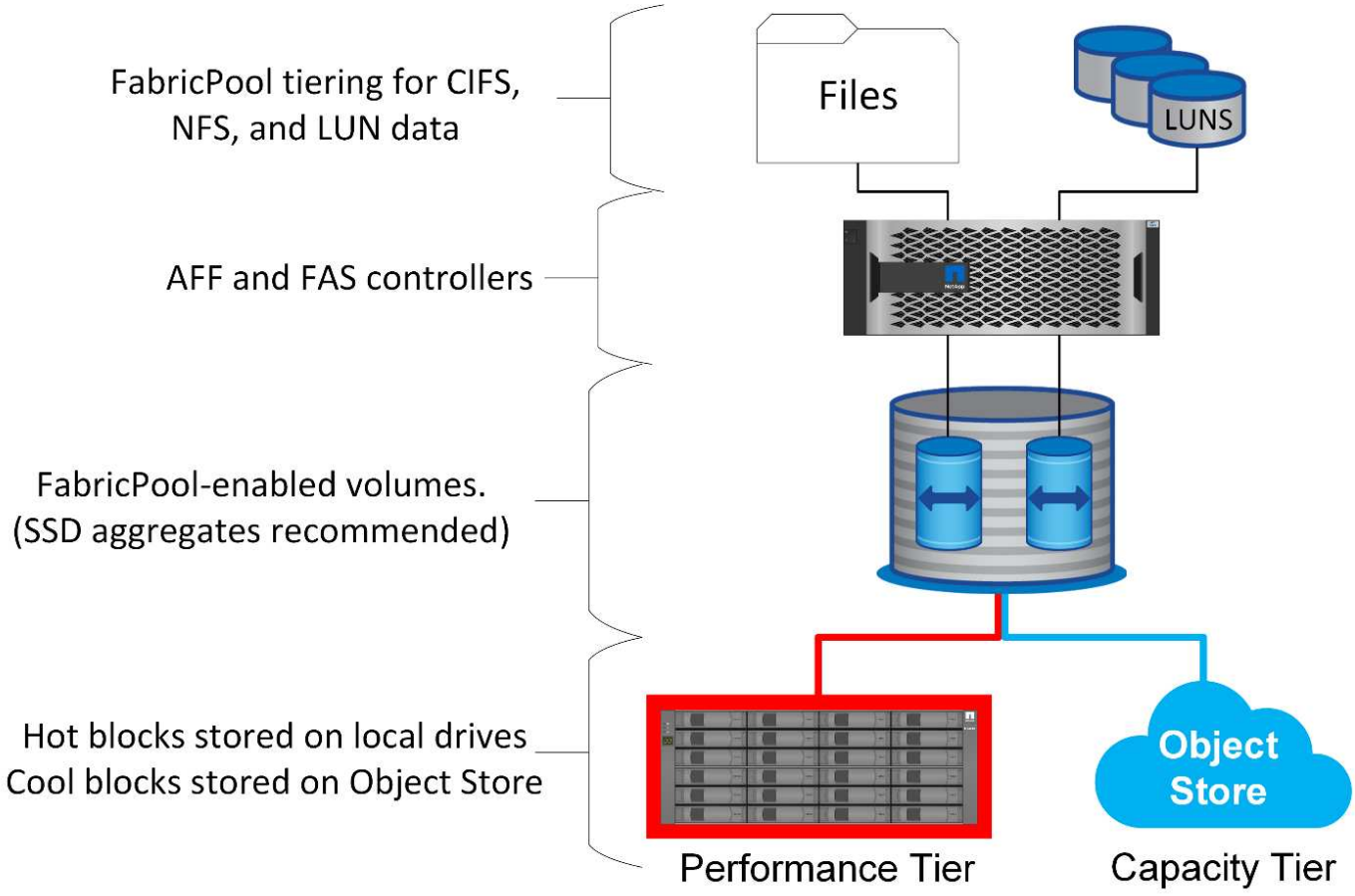
Überblick

Um zu verstehen, wie sich FabricPool Tiering auf Oracle und andere Datenbanken auswirkt, benötigen Sie ein Verständnis der Low-Level-FabricPool-Architektur.

Der Netapp Architektur Sind

FabricPool ist eine Tiering-Technologie, mit der Blöcke als „heiß“ oder „kalt“ klassifiziert werden und in dem Storage Tier platziert werden, der am besten geeignet ist. Die Performance-Tier befindet sich am häufigsten auf SSD-Storage und hostet die wichtigen Datenblöcke. Die Kapazitäts-Tier befindet sich in einem Objektspeicher und hostet die kühlen Datenblöcke. Unterstützung für Objekt-Storage: NetApp StorageGRID, ONTAP S3, Microsoft Azure Blob Storage, Alibaba Cloud Object Storage-Service, IBM Cloud Object Storage, Google Cloud Storage und Amazon AWS S3

Es stehen mehrere Tiering-Richtlinien zur Verfügung, die steuern, wie Blöcke als „heiß“ oder „kalt“ klassifiziert werden. Die Richtlinien lassen sich für einzelne Volumes festlegen und bei Bedarf ändern. Es werden nur die Datenblöcke zwischen den Performance- und Kapazitäts-Tiers verschoben. Die Metadaten, die die Struktur der LUN und des File-Systems definieren, verbleiben immer auf der Performance-Tier. Dadurch wird das Management auf ONTAP zentralisiert. Dateien und LUNs unterscheiden sich offenbar nicht von Daten, die auf einer anderen ONTAP-Konfiguration gespeichert sind. Der NetApp AFF oder FAS Controller wendet die definierten Richtlinien an, um Daten auf die entsprechende Tier zu verschieben.



Objektspeicher-Anbieter

Bei Objekt-Storage-Protokollen werden einfache HTTP- oder HTTPS-Anfragen zum Speichern einer großen Anzahl von Datenobjekten verwendet. Der Zugriff auf den Objektspeicher muss zuverlässig sein, da der Datenzugriff von ONTAP von der umgehende Erfüllung von Anfragen abhängt. Zu den Optionen gehören Amazon S3 Standard und infrequent Access sowie Microsoft Azure Hot and Cool Blob Storage, IBM Cloud und Google Cloud. Archivierungsoptionen wie Amazon Glacier und Amazon Archive werden nicht unterstützt, da die zum Abrufen von Daten erforderliche Zeit die Toleranzen der Host-Betriebssysteme und -Applikationen überschreiten kann.

Zudem wird NetApp StorageGRID unterstützt und stellt eine optimale Lösung der Enterprise-Klasse dar. Es ist ein hochperformantes, skalierbares und hochsicheres Objekt-Storage-System, das geografische Redundanz für FabricPool Daten und andere Objektspeicher-Applikationen bietet, die zunehmend Teil von Enterprise-Applikationsumgebungen sind.

StorageGRID kann zudem die Kosten senken, indem es die Egress-Gebühren vermeidet, die viele Public-Cloud-Provider beim Lesen der Daten aus ihren Services auferlegen.

Daten und Metadaten

Beachten Sie, dass der Begriff „Daten“ hier für die tatsächlichen Datenblöcke gilt, nicht für die Metadaten. Es werden nur Datenblöcke als Tiering übertragen, wobei die Metadaten in der Performance-Tier verbleiben. Darüber hinaus wird der Status eines Blocks als „heiß“ oder „kalt“ nur beeinflusst, wenn der eigentliche Datenblock gelesen wird. Das einfache Lesen des Namens, des Zeitstempels oder der Eigentümermetadaten einer Datei hat keine Auswirkung auf den Speicherort der zugrunde liegenden Datenblöcke.

Backups

Obwohl FabricPool den Storage-Platzbedarf deutlich reduzieren kann, ist es nicht für sich genommen eine Backup-Lösung. NetApp WAFL Metadaten bleiben immer auf der Performance-Tier. Falls ein schwerwiegender Ausfall die Performance-Tier zerstört, kann keine neue Umgebung aus den Daten auf der Kapazitäts-Tier erstellt werden, da sie keine WAFL-Metadaten enthält.

FabricPool kann jedoch Teil einer Backup-Strategie werden. FabricPool lässt sich beispielsweise mit der Replizierungstechnologie NetApp SnapMirror konfigurieren. Jede Hälfte der Spiegelung kann über eine eigene Verbindung mit einem Objekt-Storage-Ziel verfügen. Daraus ergeben sich zwei unabhängige Kopien der Daten. Die primäre Kopie besteht aus den Blöcken auf der Performance-Tier und den zugehörigen Blöcken auf der Kapazitäts-Tier, während das Replikat einen zweiten Satz von Performance- und Kapazitätsblöcken darstellt.

Tiering-Richtlinien

Tiering-Richtlinien

In ONTAP stehen vier Richtlinien zur Verfügung, die steuern, wie Oracle-Daten auf der Performance-Tier zu einem Kandidaten für die Verlagerung auf die Kapazitäts-Tier werden.

Nur Snapshot

Der `snapshot-only` `tiering-policy` Gilt nur für Blöcke, die nicht mit dem aktiven Dateisystem gemeinsam genutzt werden. Im Wesentlichen führt dies zum Tiering von Datenbank-Backups. Blöcke eignen sich als Tiering-Kandidaten, nachdem ein Snapshot erstellt wurde und der Block dann überschrieben wird. Das Ergebnis ist ein Block, der nur innerhalb des Snapshots vorhanden ist. Die Verzögerung vor einem `snapshot-only` Der Block wird als cool betrachtet und wird vom gesteuert `tiering-minimum-cooling-days` Einstellung für die Lautstärke. Der Bereich ab ONTAP 9.8 liegt zwischen 2 und 183 Tagen.

Viele Datensätze verfügen über niedrige Änderungsraten, wodurch diese Richtlinien nur minimal eingespart werden. Eine typische Datenbank mit ONTAP hat beispielsweise eine Änderungsrate von weniger als 5 % pro Woche. Protokolle für Datenbankarchive können umfangreichen Speicherplatz belegen, existieren jedoch normalerweise weiterhin im aktiven File-System und sind daher nicht für Tiering im Rahmen dieser Richtlinie geeignet.

Automatisch

Der `auto` die Tiering-Richtlinie erweitert das Tiering sowohl auf Snapshot-spezifische Blöcke als auch auf Blöcke innerhalb des aktiven File-Systems. Die Verzögerung, bevor ein Block als cool betrachtet wird, wird vom gesteuert `tiering-minimum-cooling-days` Einstellung für die Lautstärke. Der Bereich ab ONTAP 9.8 liegt zwischen 2 und 183 Tagen.

Dieser Ansatz ermöglicht Tiering-Optionen, die mit dem nicht verfügbar sind `snapshot-only` Richtlinie: Eine Datensicherungsrichtlinie kann beispielsweise die Aufbewahrung bestimmter Protokolldateien von 90 Tagen erfordern. Wenn Sie einen Abkühlzeitraum von 3 Tagen festlegen, werden Protokolldateien, die älter als 3 Tage sind, aus der Performance-Schicht verschoben. Dadurch wird ein erheblicher Teil des Speicherplatzes auf dem Performance-Tier freigesetzt, und Sie können die Daten der gesamten 90 Tage anzeigen und managen.

Keine

Der `none` die tiering-Richtlinie verhindert, dass zusätzliche Blöcke von der Storage-Ebene aus verschoben werden, doch alle Daten, die sich noch in der Kapazitäts-Tier befinden, bleiben bis sie gelesen werden. Wenn

der Block dann gelesen wird, wird er zurückgezogen und auf die Performance-Tier platziert.

Der Hauptgrund für die Verwendung des `none` mittels tiering-Richtlinie soll verhindert werden, dass Blöcke in Tiers verschoben werden, es könnte sich jedoch nützlich sein, die Richtlinien im Laufe der Zeit zu ändern. Nehmen wir beispielsweise an, dass ein bestimmter Datensatz häufig auf die Kapazitätsebene gestaffelt ist, doch entsteht ein unerwarteter Bedarf an vollständigen Performance-Funktionen. Die Richtlinie kann geändert werden, um ein zusätzliches Tiering zu vermeiden und sicherzustellen, dass alle Blöcke, die bei einer Zunahme der I/O-Vorgänge zurückgelesen werden, weiterhin in der Performance-Tier verbleiben.

Alle

Der `all` Die tiering-Richtlinie ersetzt die `backup` Richtlinie ab ONTAP 9.6. Der `backup` Richtlinie gilt nur für Datensicherungs-Volumes, d. h. ein Ziel für SnapMirror oder NetApp SnapVault. Der `all` Richtlinienfunktionen identisch, aber nicht beschränkt auf Datensicherungs-Volumes

Mit dieser Richtlinie gelten Blöcke sofort als „cool“ und können sofort auf die Kapazitätsebene verschoben werden.

Diese Richtlinie eignet sich besonders für langfristige Backups. Es kann auch als eine Form von Hierarchical Storage Management (HSM) verwendet werden. In der Vergangenheit wurde HSM häufig verwendet, um die Datenblöcke einer Datei auf Band zu verschieben, während die Datei selbst im Dateisystem sichtbar gehalten wurde. Ein FabricPool Volume mit dem `all` Richtlinien ermöglichen das Speichern von Dateien in einem sichtbaren und leicht zu verwaltenden System, wobei jedoch so gut wie kein Speicherplatz auf der lokalen Storage Tier belegt wird.

Abrufrichtlinien

Die Tiering-Richtlinien steuern, welche Oracle-Datenbankblöcke von der Performance-Tier auf die Kapazitäts-Tier verschoben werden. Abrufrichtlinien steuern, was passiert, wenn ein gestaffeltes Block gelesen wird.

Standard

Alle FabricPool-Volumes sind zunächst auf festgelegt `default`, D.h. das Verhalten wird durch die `Cloud-Retrieval-Policy gesteuert. `das genaue Verhalten hängt von der verwendeten Tiering Policy ab.

- `auto`- Nur zufällig gelesene Daten abrufen
- `snapshot-only`- Alle sequentiellen oder zufällig gelesenen Daten abrufen
- `none`- Alle sequentiellen oder zufällig gelesenen Daten abrufen
- `all`- Daten nicht aus der Kapazitätsebene abrufen

Gelesen

Einstellung `cloud-retrieval-policy` Das Lesen überschreibt das Standardverhalten, sodass ein Lesen von Tiered-Daten dazu führt, dass diese Daten an die Performance-Tier zurückgegeben werden.

Ein Volume könnte beispielsweise lange Zeit unter der wenig verwendet worden sein `auto` die tiering-Richtlinie, und die meisten Blöcke sind nun Tiered Storage.

Wenn bei einer unerwarteten Änderung des Geschäfts ein Teil der Daten wiederholt gescannt werden muss, um einen bestimmten Bericht zu erstellen, kann es wünschenswert sein, den zu ändern `cloud-retrieval-policy` Bis `on-read` Um sicherzustellen, dass alle gelesenen Daten in die Performance-Tier zurückgegeben

werden, einschließlich sequenzieller und zufällig gelesener Daten. Dies würde die Performance sequenzieller I/O-Vorgänge für das Volume verbessern.

Heraufstufen

Das Verhalten der „heraufstufen“-Richtlinie hängt von der Tiering-Richtlinie ab. Wenn die Tiering-Richtlinie lautet `auto`, Dann Einstellung der `cloud-retrieval-policy `to `promote` Ruft beim nächsten Tiering-Scan alle Blöcke aus der Kapazitäts-Tier zurück.

Wenn die Tiering-Richtlinie lautet `snapshot-only`, Dann sind die einzigen Blöcke, die zurückgegeben werden, die mit dem aktiven Dateisystem verbunden sind. Normalerweise hätte dies keine Auswirkung, weil die einzigen Blöcke unter das gestaffelt wären `snapshot-only` Richtlinie wären Blöcke, die ausschließlich mit Snapshots verknüpft wären. Es gäbe keine Tiered Blocks im aktiven File-System.

Wenn jedoch die Daten auf einem Volume von einem Volume-SnapRestore oder Datei-Klon-Vorgang aus einem Snapshot wiederhergestellt wurden, können einige der Blöcke, die aufgrund ihrer lediglich mit Snapshots verknüpften Speicherebenen verschoben wurden, jetzt vom aktiven File-System benötigt werden. Es kann wünschenswert sein, die vorübergehend zu ändern `cloud-retrieval-policy` Richtlinie an `promote` Alle lokal erforderlichen Blöcke schnell abrufen.

Nie

Nehmen Sie keine Blöcke aus der Kapazitäts-Tier heraus.

Tiering-Strategien

Vollständiges Datei-Tiering

FabricPool Tiering wird zwar auf Block-Ebene ausgeführt, kann jedoch in einigen Fällen für Tiering auf Dateiebene verwendet werden.

Viele Applikationsdatensätze sind nach Datum geordnet. Solche Daten sind im Allgemeinen immer seltener zugänglich, wenn sie älter werden. Beispielsweise verfügt eine Bank möglicherweise über ein Repository mit PDF-Dateien, die fünf Jahre Kundenabrechnungen enthalten, aber nur die letzten Monate sind aktiv. FabricPool kann verwendet werden, um ältere Datendateien in die Kapazitäts-Tier zu verschieben. Eine Abkühlzeit von 14 Tagen würde dafür sorgen, dass die letzten 14 Tage der PDF-Dateien auf der Performance-Ebene verbleiben. Darüber hinaus würden Dateien, die mindestens alle 14 Tage gelesen werden, „heiß“ bleiben und daher auf der Performance-Ebene verbleiben.

Richtlinien

Um einen dateibasierten Tiering-Ansatz zu implementieren, müssen Sie über Dateien verfügen, die geschrieben und nicht nachträglich geändert werden. Der `tiering-minimum-cooling-days` Richtlinien sollten so hoch eingestellt werden, dass Dateien, die Sie möglicherweise benötigen, auf der Performance-Tier verbleiben. Ein Datensatz, für den die letzten 60 Tage Daten mit optimaler Performance benötigt werden, erfordert beispielsweise die Einstellung `tiering-minimum-cooling-days` Bis 60. Ähnliche Ergebnisse lassen sich auch anhand der Dateizugriffsmuster erzielen. Wenn beispielsweise die Daten der letzten 90 Tage benötigt werden und die Applikation auf diese 90-Tage-Zeitspanne zugreift, verbleiben die Daten in der Performance-Tier. Durch Einstellen der `tiering-minimum-cooling-days` Zeitraum bis 2, erhalten Sie prompt Tiering, nachdem die Daten weniger aktiv werden.

Der `auto` Eine Richtlinie ist für das Tiering dieser Blöcke erforderlich, da nur die `auto` Die Richtlinie wirkt sich auf Blöcke aus, die sich im aktiven Filesystem befinden.

Jeder Zugriff auf Daten setzt die Heatmap-Daten zurück. Virus-Scan, Indizierung und sogar Backup-Aktivitäten, die die Quelldateien lesen, verhindern Tiering, da dies erforderlich ist `tiering-minimum-cooling-days` Schwellenwert wird nie erreicht.

Tiering von partiellen Dateien

Da FabricPool auf Block-Ebene arbeitet, können geänderte Dateien teilweise auf Objekt-Storage verschoben werden und dennoch nur teilweise auf Performance-Tier verbleiben.

Dies ist bei Datenbanken üblich. Datenbanken, für die bekanntermaßen inaktive Blöcke enthalten sind, eignen sich auch für das FabricPool Tiering. Beispielsweise kann eine Supply-Chain-Management-Datenbank historische Informationen enthalten, die bei Bedarf verfügbar sein müssen, aber während des normalen Betriebs nicht aufgerufen werden. Mit FabricPool können die inaktiven Blöcke selektiv verschoben werden.

Beispielsweise Datendateien, die auf einem FabricPool Volume mit einem ausgeführt werden `tiering-minimum-cooling-days` Im Zeitraum von 90 Tagen werden sämtliche Blöcke aufbewahrt, auf die in den vorangegangenen 90 Tagen auf der Performance Tier zugegriffen wurde. Alle Daten, auf die 90 Tage lang nicht zugegriffen wird, werden jedoch auf die Kapazitäts-Tier verlagert. In anderen Fällen bleiben bei normalen Applikationsaktivitäten die richtigen Blöcke auf der richtigen Tier erhalten. Wenn beispielsweise eine Datenbank normalerweise dazu verwendet wird, die Daten der letzten 60 Tage regelmäßig zu verarbeiten, ist dies wesentlich geringer `tiering-minimum-cooling-days` Zeitraum kann festgelegt werden, da die natürliche Aktivität der Anwendung dafür sorgt, dass Blöcke nicht vorzeitig verschoben werden.

Der `auto` Richtlinien sollten mit Vorsicht bei Datenbanken verwendet werden. Viele Datenbanken verfügen über periodische Aktivitäten wie etwa Vorgänge zum Quartalsende oder die Neuindizierung. Wenn der Zeitraum dieser Vorgänge größer ist als der `tiering-minimum-cooling-days` Es können Performance-Probleme auftreten. Wenn zum Quartalsende beispielsweise 1 TB an Daten verarbeitet werden müssen, die ansonsten nicht verarbeitet wurden, befinden sich diese Daten möglicherweise nun auf der Kapazitäts-Tier. Lesezugriffe von der Kapazitäts-Tier sind oft extrem schnell und verursachen möglicherweise keine Performance-Probleme. Die genauen Ergebnisse hängen jedoch von der Objektspeicher-Konfiguration ab.

Richtlinien

Der `tiering-minimum-cooling-days` Die Richtlinie sollte so hoch eingestellt werden, dass Dateien, die auf der Performance-Tier erforderlich sind, aufbewahrt werden. Beispielsweise müsste eine Datenbank, in der die letzten 60 Tage Daten bei einer optimalen Performance benötigt werden, die festlegen `tiering-minimum-cooling-days` Zeitraum bis 60 Tage. Ähnliche Ergebnisse lassen sich auch anhand der Zugriffsmuster von Dateien erzielen. Wenn beispielsweise die Daten der letzten 90 Tage benötigt werden und die Applikation auf diese 90-Tage-Datenspanne zugreift, verbleiben die Daten in der Performance-Tier. Einstellen des `tiering-minimum-cooling-days` Zeitraum bis 2 Tage würde die Daten sofort nach dem Zeitpunkt verschieben, an dem die Daten weniger aktiv sind.

Der `auto` Eine Richtlinie ist für das Tiering dieser Blöcke erforderlich, da nur die `auto` Die Richtlinie wirkt sich auf Blöcke aus, die sich im aktiven Filesystem befinden.

Jeder Zugriff auf Daten setzt die Heatmap-Daten zurück. Daher verhindert die Überprüfung der vollständigen Tabelle der Datenbank und sogar die Backup-Aktivitäten, die die Quelldateien lesen, Tiering, da die erforderlichen `tiering-minimum-cooling-days` Schwellenwert wird nie erreicht.

Tiering von Archivprotokollen

Die wahrscheinlich wichtigste Verwendung für FabricPool ist die Verbesserung der Effizienz bekannter, kalter Daten, wie z. B. Transaktions-Logs der Datenbank.

Die meisten relationalen Datenbanken arbeiten im Transaktionsprotokoll-Archivierungsmodus, um Point-in-Time Recovery bereitzustellen. Änderungen an den Datenbanken werden durch die Aufzeichnung der Änderungen in den Transaktionsprotokollen vorgenommen und das Transaktionsprotokoll wird ohne Überschreibung beibehalten. Dies kann zur Anforderung führen, eine enorme Menge an archivierten Transaktions-Logs aufzubewahren. Ähnliche Beispiele gibt es bei vielen anderen Applikations-Workflows, die Daten generieren, die aufbewahrt werden müssen, auf die jedoch mit hoher Wahrscheinlichkeit jemals zugegriffen werden wird.

FabricPool löst diese Probleme mit einer einzigen Lösung mit integriertem Tiering. Dateien werden gespeichert und bleiben an ihrem üblichen Speicherort zugänglich, belegen jedoch praktisch keinen Speicherplatz auf dem primären Array.

Richtlinien

Verwenden Sie A `tiering-minimum-cooling-days` Eine Richtlinie von wenigen Tagen führt zur Aufbewahrung von Blöcken in den kürzlich erstellten Dateien (die Dateien sind, die in naher Zukunft am wahrscheinlichsten erforderlich sind) auf der Performance-Tier. Die Datenblöcke aus älteren Dateien werden dann auf die Kapazitäts-Tier verschoben.

Der `auto` Erzwingt sofortiges Tiering, wenn der Kühlschwellenwert erreicht wurde, unabhängig davon, ob die Protokolle gelöscht wurden oder weiterhin im primären Dateisystem vorhanden sind. Auch das Speichern aller potenziell erforderlichen Protokolle an einer zentralen Stelle im aktiven Filesystem vereinfacht das Management. Es gibt keinen Grund, Snapshots zu durchsuchen, um eine Datei zu finden, die wiederhergestellt werden muss.

Einige Applikationen, wie z. B. Microsoft SQL Server, schneiden Transaktions-Log-Dateien während von Backup-Vorgängen ab, sodass sich die Protokolle nicht mehr im aktiven File-System befinden. Die Kapazität kann mithilfe des gespeichert werden `snapshot-only` tiering-Richtlinie, aber die `auto` Die Richtlinie ist für Protokolldaten nicht nützlich, da Protokolldaten im aktiven Dateisystem selten abgekühlt werden sollten.

Snapshot Tiering

Die erste Version von FabricPool war auf den Backup-Anwendungsfall ausgerichtet. Als einzige Art von Blöcken, die Tiering ermöglichen konnten, handelte es sich um Blöcke, die nicht mehr mit Daten im aktiven File-System verknüpft waren. Daher können nur die Snapshot Datenblöcke auf diese Kapazitäts-Tier verschoben werden. Dies bleibt eine der sichersten Tiering-Optionen, wenn Sie sicherstellen müssen, dass die Performance nie beeinträchtigt wird.

Richtlinien: Lokale Snapshots

Es gibt zwei Optionen für das Tiering inaktiver Snapshot-Blöcke auf die Kapazitäts-Tier. Zunächst einmal die `snapshot-only` Die Richtlinie zielt nur auf die Snapshot-Blöcke ab. Obwohl der `auto` Die Richtlinie umfasst die `snapshot-only` Blöcke, sondern auch Tiering Blöcke aus dem aktiven File-System. Dies ist möglicherweise nicht wünschenswert.

Der `tiering-minimum-cooling-days` Der Wert sollte auf einen Zeitraum festgelegt werden, in dem Daten, die während einer Wiederherstellung erforderlich sein könnten, auf der Performance-Tier zur Verfügung

stehen. So enthalten die meisten Wiederherstellungsszenarien einer kritischen Produktionsdatenbank zu einem bestimmten Zeitpunkt in den letzten Tagen einen Wiederherstellungspunkt. Einstellung A `tiering-minimum-cooling-days` Mit dem Wert 3 würde sichergestellt, dass bei einer Wiederherstellung der Datei eine Datei entsteht, die sofort die maximale Performance liefert. Alle Blöcke in den aktiven Dateien befinden sich immer noch auf schnellem Storage, ohne dass eine Wiederherstellung aus dem Kapazitäts-Tier erforderlich ist.

Richtlinien – replizierte Snapshots

Ein Snapshot, der mit SnapMirror oder SnapVault repliziert wird, der nur für die Wiederherstellung verwendet wird, sollte im Allgemeinen die FabricPool verwenden `all` Richtlinie: Bei dieser Richtlinie werden Metadaten repliziert. Alle Datenblöcke werden jedoch sofort an die Kapazitäts-Tier gesendet, was maximale Performance liefert. Die meisten Recovery-Prozesse arbeiten mit sequenziellem I/O, was von vornherein effizient ist. Die Recovery-Zeit vom Zielort des Objektspeichers ist zu bewerten, in einer gut durchdachten Architektur muss dieser Recovery-Prozess jedoch nicht wesentlich langsamer sein als die Wiederherstellung von lokalen Daten.

Wenn die replizierten Daten auch für das Klonen verwendet werden sollen, wird der verwendet `auto` Die Politik ist angemessener, mit einem `tiering-minimum-cooling-days` Wert, der Daten umfasst, von denen erwartet wird, dass sie regelmäßig in einer Klonumgebung verwendet werden. Der aktive Arbeitsdatensatz einer Datenbank kann beispielsweise Daten enthalten, die in den letzten drei Tagen gelesen oder geschrieben wurden, aber es können auch weitere Verlaufsdaten von 6 Monaten enthalten sein. Wenn ja, dann die `auto` Durch eine Richtlinie am Ziel von SnapMirror wird das Arbeitsdatensatz auf der Performance-Ebene verfügbar.

Backup-Tiering

Zu den herkömmlichen Applikations-Backups gehören Produkte wie der Oracle Recovery Manager, die dateibasierte Backups außerhalb des Standorts der Originaldatenbank erstellen.

```
 `tiering-minimum-cooling-days` policy of a few days preserves the most
recent backups, and therefore the backups most likely to be required for
an urgent recovery situation, on the performance tier. The data blocks of
the older files are then moved to the capacity tier.
Der `auto` Die Richtlinie ist die am besten geeignete Richtlinie für
Backup-Daten. Dadurch wird ein sofortiges Tiering sichergestellt, wenn der
Kühlschwellenwert erreicht wurde, unabhängig davon, ob die Dateien
gelöscht wurden oder weiterhin im primären Dateisystem vorhanden sind. Das
Speichern aller potenziell erforderlichen Dateien an einem zentralen
Speicherort im aktiven Dateisystem vereinfacht ebenfalls das Management.
Es gibt keinen Grund, Snapshots zu durchsuchen, um eine Datei zu finden,
die wiederhergestellt werden muss.
```

Der `snapshot-only` Richtlinien können zwar funktionieren, sie gelten jedoch nur für Blöcke, die sich nicht mehr im aktiven File-System befinden. Daher müssen Dateien auf einer NFS- oder SMB-Freigabe vor dem Daten-Tiering zuerst gelöscht werden.

Diese Richtlinie wäre bei einer LUN-Konfiguration sogar noch weniger effizient, da beim Löschen einer Datei aus einer LUN nur Dateiverweise aus den Metadaten des Filesystems entfernt werden. Die tatsächlichen Blöcke auf den LUNs bleiben vorhanden, bis sie überschrieben werden. Dies kann zu einer sehr langen Verzögerung zwischen dem Löschen einer Datei und dem Überschreiben der Blöcke führen und zu Tiering-

Kandidaten werden. Der Wechsel des bietet einige Vorteile `snapshot-only` Blöcke auf die Kapazitäts-Tier, aber insgesamt funktioniert das FabricPool Management von Backup-Daten am besten mit der `auto` Richtlinie:

Mit diesem Ansatz können Benutzer den für Backups erforderlichen Speicherplatz effizienter managen. FabricPool selbst ist jedoch keine Backup-Technologie. Das Tiering von Backup-Dateien in Objektspeicher vereinfacht das Management, da die Dateien noch auf dem ursprünglichen Storage-System sichtbar sind, die Datenblöcke im Zielspeicherort jedoch vom ursprünglichen Storage-System abhängig sind. Wenn das Quell-Volume verloren geht, sind die Objektspeicher-Daten nicht mehr nutzbar.

Unterbrechungen des Zugriffs auf Objektspeicher

Tiering ein Datensatz mit FabricPool ergibt eine Abhängigkeit zwischen dem primären Storage Array und der Objektspeicher-Ebene. Es gibt zahlreiche Objektspeicher-Optionen, die eine unterschiedliche Verfügbarkeit bieten. Es ist wichtig, die Auswirkungen eines möglichen Verbindungsverlusts zwischen dem primären Storage-Array und dem Objekt-Storage-Tier zu verstehen.

Wenn ein an ONTAP ausgehändigt I/O Daten aus der Kapazitäts-Tier benötigt und ONTAP die Kapazitäts-Tier nicht erreichen kann, um Blöcke abzurufen, wird schließlich ein Ausfall des I/O-Systems erreicht. Die Auswirkung dieses Timeouts hängt vom verwendeten Protokoll ab. In einer NFS-Umgebung antwortet ONTAP je nach Protokoll entweder mit einer EJUKEBOX- oder EDELAY-Antwort. Einige ältere Betriebssysteme interpretieren dies möglicherweise als Fehler, aber aktuelle Betriebssysteme und aktuelle Patch-Level des Oracle Direct NFS-Clients behandeln dies als Retrievable-Fehler und warten weiterhin auf den Abschluss des I/O.

Ein kürzeres Timeout gilt für SAN-Umgebungen. Wenn ein Block in der Objektspeicherumgebung erforderlich ist und zwei Minuten lang nicht erreichbar bleibt, wird ein Lesefehler an den Host zurückgegeben. Das ONTAP Volume und die LUNs bleiben online, das Host-Betriebssystem kennzeichnet das Filesystem jedoch möglicherweise als fehlerhaft.

Konnektivitätsprobleme bei Objekt-Storage `snapshot-only` Die Richtlinie ist weniger bedenklich, da nur Backup-Daten als Tiering übertragen werden. Kommunikationsprobleme würden die Datenwiederherstellung verlangsamen, würden jedoch die aktive Nutzung der Daten nicht beeinträchtigen. Der `auto` Und `all` Mithilfe von Richtlinien wird das Tiering „kalter“ Daten von der aktiven LUN ermöglicht. Ein Fehler beim Abrufen von Objektspeicher-Daten kann sich somit auf die Datenbankverfügbarkeit auswirken. Eine SAN-Implementierung mit diesen Richtlinien sollte nur mit Objektspeicher der Enterprise-Klasse und Netzwerkverbindungen genutzt werden, die auf Hochverfügbarkeit ausgelegt sind. NetApp StorageGRID ist die überlegene Option.

Oracle Datensicherung

Datensicherung mit ONTAP

NetApp weiß, dass die geschäftskritischsten Daten in Datenbanken zu finden sind.

Ein Unternehmen kann nicht ohne Zugriff auf seine Daten arbeiten, und manchmal definieren die Daten das Unternehmen. Diese Daten müssen geschützt werden. Bei der Datensicherung geht es jedoch mehr als nur um das Sicherstellen eines nutzbaren Backups. Es geht darum, die Backups schnell und zuverlässig durchzuführen und diese sicher zu speichern.

Die andere Seite der Datensicherung ist die Datenwiederherstellung. Wenn auf Daten nicht zugegriffen werden kann, ist das Unternehmen betroffen und kann nicht mehr in Betrieb sein, bis die Daten wiederhergestellt werden. Dieser Prozess muss schnell und zuverlässig sein. Schließlich müssen die meisten Datenbanken vor Ausfällen geschützt werden, was bedeutet, dass ein Replikat der Datenbank beibehalten wird. Das Replikat muss ausreichend aktuell sein. Außerdem muss es schnell und einfach sein, das Replikat zu einer voll funktionsfähigen Datenbank zu machen.

Diese Dokumentation ersetzt den zuvor veröffentlichten technischen Bericht *TR-4591: Oracle Data Protection: Backup, Recovery und Replication.*

Planung

Die richtige Datensicherungsarchitektur hängt von den geschäftlichen Anforderungen für die Datenaufbewahrung, Recovery-Fähigkeit und Ausfalltoleranz bei verschiedenen Ereignissen ab.

Betrachten Sie beispielsweise die Anzahl der im Umfang enthaltenen Applikationen, Datenbanken und wichtigen Datensätze. Die Entwicklung einer Backup-Strategie für einen einzelnen Datensatz gewährleistet, dass die Compliance mit typischen SLAs relativ unkompliziert ist, da nicht viele Objekte zu managen sind. Je mehr Datensätze es gibt, desto komplizierter wird das Monitoring und Administratoren müssen sich zunehmend mit dem Vermeiden von Backup-Fehlern befassen. Wenn eine Umgebung also die Skalierung von Cloud- und Service-Provider-Umgebungen erreicht, braucht es einen ganz anderen Ansatz.

Die Datensatzgröße wirkt sich auch auf die Strategie aus. Es gibt beispielsweise viele Optionen für Backup und Recovery mit einer Datenbank mit 100 GB, da die Datenmenge so klein ist. Das einfache Kopieren der Daten von Backup-Medien mit herkömmlichen Tools bietet normalerweise eine ausreichende RTO für die Recovery. Eine 100-TB-Datenbank benötigt normalerweise eine komplett andere Strategie, es sei denn, das RTO erlaubt einen mehrtägigen Ausfall. In diesem Fall könnte ein herkömmliches Backup und Recovery auf Basis von Kopien akzeptabel sein.

Schließlich gibt es Faktoren, die nicht dem Backup- und Recovery-Prozess selbst unterliegen. Gibt es zum Beispiel Datenbanken, die kritische Produktionsaktivitäten unterstützen, was das Recovery zu einem seltenen Ereignis macht, das nur von erfahrenen DBAs durchgeführt wird? Sind Datenbanken alternativ Teil einer großen Entwicklungsumgebung, in der häufig ein Recovery erfolgt und von einem IT-Team mit Generalisten gemanagt wird?

RTO, RPO und SLA-Planung

Mit ONTAP können Sie in einfacher Weise eine Datensicherungsstrategie für Oracle Database an Ihre Geschäftsanforderungen anpassen.

Zu diesen Anforderungen gehören Faktoren wie die Geschwindigkeit der Recovery, der maximal zulässige Datenverlust und die Anforderungen an die Aufbewahrung von Backups. Der Datensicherungsplan muss zudem verschiedene gesetzliche Vorgaben für die Datenaufbewahrung und -Wiederherstellung berücksichtigen. Schließlich müssen verschiedene Datenwiederherstellungsszenarien in Betracht gezogen werden, von der typischen und vorhersehbaren Wiederherstellung aufgrund von Benutzer- oder Applikationsfehlern bis hin zu Disaster Recovery-Szenarien, die den vollständigen Ausfall eines Standorts beinhalten.

Kleine Änderungen an Richtlinien zur Datensicherung und Wiederherstellung können sich erheblich auf die Gesamtarchitektur von Storage, Backup und Recovery auswirken. Es ist wichtig, Standards zu definieren und zu dokumentieren, bevor mit dem Design begonnen wird, um eine Verkomplizierung einer Datensicherungsarchitektur zu vermeiden. Unnötige Schutzfunktionen oder -Ebenen führen zu unnötigen Kosten und Management-Overhead. Eine zunächst übersehene Anforderung kann ein Projekt in die falsche Richtung führen oder kurzfristig Designänderungen erfordern.

Recovery-Zeitvorgabe

Die Recovery-Zeitvorgabe (Recovery Time Objective, RTO) definiert die maximal zulässige Zeit für die Recovery eines Services. Eine Personaldatenbank könnte beispielsweise eine RTO von 24 Stunden haben, da, obwohl es sehr unpraktisch wäre, den Zugriff auf diese Daten während der Arbeitszeit zu verlieren, das Unternehmen dennoch arbeiten kann. Im Gegensatz dazu würde bei einer Datenbank, die das Hauptbuch einer Bank unterstützt, eine RTO in Minuten oder sogar Sekunden gemessen werden. Ein RTO von null ist nicht möglich, da es eine Möglichkeit geben muss, zwischen einem tatsächlichen Serviceausfall und einem Routineereignis wie einem verlorenen Netzwerkpaket zu unterscheiden. Typische Anforderungen sind jedoch ein RTO von nahezu null.

Recovery-Zeitpunkt

Der Recovery Point Objective (RPO) definiert den maximal tolerierbaren Datenverlust. In vielen Fällen wird der RPO lediglich durch die Häufigkeit von Snapshots oder snapmirror Updates bestimmt.

In manchen Fällen lässt sich der RPO-Wert aggressiver einsetzen, da er bestimmte Daten selektiv häufiger schützt. Im Datenbankkontext ist der RPO in der Regel eine Frage, wie viele Protokolldaten in einer bestimmten Situation verloren gehen können. In einem typischen Recovery-Szenario, bei dem eine Datenbank aufgrund eines Produktfehlers oder eines Benutzerfehlers beschädigt wird, sollte der RPO gleich null sein, d. h. es darf keine Daten verloren gehen. Bei der Wiederherstellung wird eine frühere Kopie der Datenbankdateien wiederhergestellt und anschließend die Protokolldateien wiedergegeben, um den Datenbankstatus auf den gewünschten Zeitpunkt zu bringen. Die für diesen Vorgang erforderlichen Protokolldateien sollten sich bereits am ursprünglichen Speicherort befinden.

In ungewöhnlichen Szenarien können Protokolldaten verloren gehen. Zum Beispiel eine versehentliche oder böswillige `rm -rf *` Der Datenbankdateien können zum Löschen aller Daten führen. Die einzige Option wäre die Wiederherstellung aus dem Backup, einschließlich Protokolldateien, und einige Daten würden unweigerlich verloren gehen. Die einzige Option zur Verbesserung des RPO in einer herkömmlichen Backup-Umgebung besteht in der Durchführung wiederholter Backups der Protokolldaten. Dies hat jedoch Einschränkungen aufgrund der ständigen Datenverschiebung und der Schwierigkeiten, ein Backup-System als ständig laufenden Service zu warten. Einer der Vorteile erweiterter Storage-Systeme besteht in der Möglichkeit, Daten vor versehentlichen oder böswilligen Schäden an Dateien zu schützen und somit ein besseres RPO ohne Datenverschiebung zu ermöglichen.

Disaster Recovery

Disaster Recovery umfasst die IT-Architektur, Richtlinien und Verfahren, die zur Wiederherstellung eines Services bei einem physischen Ausfall erforderlich sind. Dies kann Überschwemmungen, Brände oder Personen sein, die mit böswilliger oder fahrlässiger Absicht handeln.

Disaster Recovery ist mehr als nur eine Reihe von Recovery-Verfahren. Der gesamte Prozess umfasst die Identifizierung der verschiedenen Risiken, die Definition der Anforderungen an die Datenwiederherstellung und die Servicekontinuität sowie die Bereitstellung der richtigen Architektur mit den zugehörigen Verfahren.

Bei der Festlegung von Datensicherungsanforderungen ist es entscheidend, zwischen den typischen RPO- und RTO-Anforderungen und den für die Disaster Recovery erforderlichen RPO- und RTO-Anforderungen zu unterscheiden. Einige Applikationsumgebungen erfordern einen RPO von null und ein RTO von nahezu null für Datenverluste – von einem relativ normalen Benutzerfehler bis hin zu einem Brand, der ein Datacenter zerstört. Für diese hohen Schutzniveaus gibt es jedoch Kosten- und administrative Konsequenzen.

Im Allgemeinen sollten die Anforderungen an die nicht-Disaster-Recovery aus zwei Gründen strikt erfüllt werden. Zunächst sind Anwendungsfehler und Benutzerfehler, die zu Datenschäden führen, bis zu dem Punkt vorhersehbar, an dem sie fast unvermeidlich sind. Zweitens ist es nicht schwierig, eine Backup-Strategie zu entwickeln, die einen RPO von null und ein RTO von niedrigen Vorgaben liefern kann, solange das Storage-

System nicht zerstört wird. Es gibt keinen Grund, ein erhebliches Risiko, das leicht behoben werden kann, nicht anzugehen. Deshalb sollten die RPO- und RTO-Ziele für die lokale Recovery aggressiv sein.

Disaster Recovery-RTO- und RPO-Anforderungen variieren stärker, je nach Wahrscheinlichkeit eines Ausfalls und den Folgen des damit verbundenen Datenverlusts oder der Unterbrechung des Geschäftsbetriebs. RPO- und RTO-Anforderungen sollten auf den tatsächlichen geschäftlichen Anforderungen basieren und nicht auf allgemeinen Prinzipien. Sie müssen mehrere logische und physische Ausfallszenarien berücksichtigen.

Logische Ausfälle

Zu logischen Katastrophen gehören Datenbeschädigungen durch Benutzer, Applikations- oder Betriebssystemfehler und Fehlfunktionen. Zu logischen Katastrophen können auch böswillige Angriffe durch externe Parteien mit Viren oder Würmern gehören oder die Ausnutzung von Schwachstellen von Applikationen. In diesen Fällen wird die physische Infrastruktur unbeschädigt, die zugrunde liegenden Daten sind jedoch nicht mehr gültig.

Eine immer häufiger vorauftretende logische Katastrophe wird als Ransomware bezeichnet. Bei ihr werden Daten mit einem Angriffsvektor verschlüsselt. Die Verschlüsselung schädigt die Daten nicht, macht sie jedoch erst verfügbar, wenn die Zahlung an einen Dritten erfolgt. Immer mehr Unternehmen sind gezielt auf Ransomware-Hacks ausgerichtet. Für diese Bedrohung bietet NetApp manipulationssichere Snapshots, bei denen nicht einmal der Storage-Administrator geschützte Daten vor dem konfigurierten Ablaufdatum ändern kann.

Physische Ausfälle

Zu physischen Ausfällen gehört der Ausfall von Komponenten einer Infrastruktur, die die Redundanzmerkmale übertreffen und zu einem Datenverlust oder erweitertem Service-Verlust führen. Der RAID-Schutz bietet beispielsweise Redundanz für Laufwerke, und die Verwendung von HBAs bietet Redundanz für FC-Port und FC-Kabel. Hardwareausfälle solcher Komponenten sind vorhersehbar und beeinträchtigen nicht die Verfügbarkeit.

In einer Unternehmensumgebung ist es in der Regel möglich, die Infrastruktur eines gesamten Standorts mit redundanten Komponenten so weit zu schützen, dass das einzige vorhersehbare physische Ausfallszenario ein vollständiger Verlust des Standorts ist. Die Planung des Disaster Recovery hängt dann von der Site-to-Site-Replizierung ab.

Synchrone und asynchrone Datensicherung

Im Idealfall würden alle Daten zwischen geografisch verteilten Standorten synchron repliziert werden. Eine solche Replikation ist nicht immer möglich oder sogar aus mehreren Gründen möglich:

- Die synchrone Replikation erhöht zwangsläufig die Schreiblatenz, da alle Änderungen an beiden Standorten repliziert werden müssen, bevor die Applikation/Datenbank mit der Verarbeitung fortfahren kann. Der daraus resultierende Performance-Effekt ist manchmal nicht akzeptabel, sodass die Verwendung von synchroner Spiegelung ausgeschlossen wird.
- Die zunehmende Einführung von 100 % SSD-Storage bedeutet, dass zusätzliche Schreiblatenz mit größerer Wahrscheinlichkeit zu verzeichnen ist, da die Performance-Erwartungen Hunderttausende IOPS und eine Latenz von unter einer Millisekunde umfassen. Um das volle Potenzial von 100 % SSDs auszuschöpfen, kann ein erneuter Besuch der Disaster-Recovery-Strategie erforderlich sein.
- Die Anzahl der Datensätze nimmt weiterhin an Byte zu. Dies stellt Unternehmen vor Herausforderungen, wenn es darum geht, genügend Bandbreite für eine synchrone Replizierung sicherzustellen.
- Die Komplexität der Datensätze nimmt zu und führt zu Herausforderungen beim Management einer umfassenden synchronen Replizierung.

- Cloud-basierte Strategien sind häufig mit höheren Replizierungsentfernungen und Latenz verbunden, wodurch die Nutzung einer synchronen Spiegelung weiterhin ausgeschlossen wird.

NetApp bietet Lösungen, die sowohl synchrone Replikation für höchste Anforderungen an die Datenwiederherstellung als auch asynchrone Lösungen für eine bessere Performance und Flexibilität beinhalten. Darüber hinaus lässt sich die NetApp Technologie nahtlos in viele Replizierungslösungen von Drittanbietern integrieren, wie z. B. Oracle DataGuard

Aufbewahrungszeit

Der letzte Aspekt einer Datensicherungsstrategie ist die Zeit für die Datenaufbewahrung, die sehr unterschiedlich sein kann.

- Eine typische Anforderung sind nächtliche Backups von 14 Tagen auf dem primären Standort und 90 Tage Backups auf einem sekundären Standort.
- Viele Kunden erstellen vierteljährliche eigenständige Archive, die auf unterschiedlichen Medien gespeichert sind.
- Eine ständig aktualisierte Datenbank benötigt möglicherweise keine Verlaufsdaten, und Backups müssen nur für einige Tage aufbewahrt werden.
- Gesetzliche Vorschriften erfordern möglicherweise die Wiederherstellbarkeit bis zu einem beliebigen Zeitpunkt jeder beliebigen Transaktion innerhalb eines Zeitfensters von 365 Tagen.

Datenbankverfügbarkeit

ONTAP wurde für eine maximale Verfügbarkeit von Oracle-Datenbanken konzipiert. Eine vollständige Beschreibung der Hochverfügbarkeitsfunktionen von ONTAP übersteigt den Rahmen dieses Dokuments. Wie bei der Datensicherheit ist jedoch ein grundlegendes Verständnis dieser Funktionalität bei der Entwicklung einer Datenbankinfrastruktur wichtig.

HA-Paare

Die Basiseinheit der Hochverfügbarkeit ist das HA-Paar. Jedes Paar enthält redundante Links, um die Replikation von Daten in NVRAM zu unterstützen. NVRAM ist kein Schreib-Cache. Der RAM im Controller dient als Schreib-Cache. Der Zweck von NVRAM besteht darin, Daten vorübergehend zu protokollieren, um Schutz vor unerwarteten Systemausfällen zu bieten. In dieser Hinsicht ähnelt es einem Datenbank-Redo-Protokoll.

Sowohl NVRAM als auch ein Datenbank-Wiederherstellungsprotokoll werden verwendet, um Daten schnell zu speichern, sodass Datenänderungen so schnell wie möglich vorgenommen werden können. Die Aktualisierung der persistenten Daten auf Laufwerken (oder Datendateien) findet erst später bei einem Prozess statt, der sowohl auf ONTAP- als auch auf den meisten Datenbankplattformen als Checkpoint bezeichnet wird. Weder NVRAM-Daten noch Datenbank-Wiederherstellungsprotokolle werden im normalen Betrieb gelesen.

Wenn ein Controller abrupt ausfällt, sind in NVRAM wahrscheinlich noch nicht gespeicherte Änderungen zu erwarten, die noch nicht auf die Laufwerke geschrieben wurden. Der Partner-Controller erkennt den Ausfall, übernimmt die Kontrolle über die Laufwerke und wendet die erforderlichen Änderungen an, die im NVRAM gespeichert wurden.

Takeover und Giveback

Takeover und Giveback beziehen sich auf den Prozess, bei dem die Verantwortung für Storage-Ressourcen zwischen Nodes in einem HA-Paar übertragen wird. Takeover und Giveback sind zweierlei Aspekte:

- Verwaltung der Netzwerkverbindung, die den Zugriff auf die Laufwerke ermöglicht
- Verwaltung der Antriebe selbst.

Die Netzwerkschnittstellen, die CIFS- und NFS-Datenverkehr unterstützen, werden sowohl mit dem Home-Standort als auch mit dem Failover-Standort konfiguriert. Eine Übernahme umfasst das Verschieben der Netzwerkschnittstellen zu ihrem temporären Home-Standort auf einer physischen Schnittstelle, die sich in denselben Subnetzen befindet wie der ursprüngliche Standort. Bei einem Giveback werden die Netzwerkschnittstellen zurück an ihre ursprünglichen Standorte verschoben. Das genaue Verhalten kann nach Bedarf angepasst werden.

Netzwerkschnittstellen, die SAN-Blockprotokolle wie iSCSI und FC unterstützen, werden während des Takeover und Giveback nicht verlagert. Stattdessen sollten LUNs mit Pfaden bereitgestellt werden, die ein vollständiges HA-Paar enthalten, was zu einem primären Pfad und einem sekundären Pfad führt.

Zusätzliche Pfade zu zusätzlichen Controllern können auch konfiguriert werden, um das Verschieben von Daten zwischen Nodes in einem größeren Cluster zu unterstützen. Dies ist jedoch nicht Teil des HA-Prozesses.

Der zweite Aspekt von Takeover und Giveback ist die Übertragung der Eigentumsrechte an den Festplatten. Der genaue Prozess hängt von mehreren Faktoren ab, einschließlich dem Grund für das Takeover/Giveback und den ausgegebenen Befehlszeilenoptionen. Das Ziel ist es, die Operation so effizient wie möglich durchzuführen. Obwohl der Gesamtprozess möglicherweise mehrere Minuten in Anspruch nimmt, kann der tatsächliche Zeitpunkt, in dem die Eigentumsrechte an dem Laufwerk von einem Node auf einen Node übertragen werden, in der Regel in Sekunden gemessen werden.

Takeover-Zeit

Host I/O durchläuft zwar eine kurze I/O-Pause bei Takeover- und Giveback-Vorgängen, jedoch sollte in einer korrekt konfigurierten Umgebung keine Applikationsunterbrechung auftreten. Der eigentliche Transitionsprozess, bei dem I/O verzögert wird, wird in der Regel in Sekunden gemessen. Der Host benötigt jedoch möglicherweise zusätzliche Zeit, um die Änderung der Datenpfade zu erkennen und die I/O-Vorgänge erneut auszuführen.

Die Art der Störung hängt vom Protokoll ab:

- Eine Netzwerkschnittstelle, die NFS- und CIFS-Datenverkehr unterstützt, stellt nach dem Übergang zu einem neuen physischen Standort eine ARP-Anforderung (Address Resolution Protocol) an das Netzwerk aus. Dies führt dazu, dass die Netzwerk-Switches ihre MAC-Adresstabellen (Media Access Control) aktualisieren und die I/O-Verarbeitung fortsetzen Im Falle von geplanten Takeover und Giveback werden Störungen in der Regel in Sekunden gemessen und oftmals nicht feststellbar. Einige Netzwerke sind möglicherweise langsamer, um die Änderung des Netzwerkpfads vollständig zu erkennen, und einige Betriebssysteme können in sehr kurzer Zeit viele I/O-Vorgänge in Warteschlange stellen, die erneut versucht werden müssen. Dadurch kann die für die I/O-Wiederaufnahme erforderliche Zeit verlängert werden
- Eine Netzwerkschnittstelle, die SAN-Protokolle unterstützt, kann nicht an einen neuen Speicherort verschoben werden. Ein Host-Betriebssystem muss den oder die verwendeten Pfade ändern. Die vom Host beobachtete I/O-Pause hängt von mehreren Faktoren ab. Aus Sicht des Storage-Systems beträgt der Zeitraum, in dem I/O nicht mehr ausgeführt werden kann, nur wenige Sekunden. Verschiedene Host-Betriebssysteme erfordern jedoch möglicherweise eine zusätzliche Zeit, damit eine I/O-Dauer vor einem

erneuten Versuch wieder aberkannt wird. Neuere Betriebssysteme können eine Pfadänderung viel schneller erkennen, aber ältere Betriebssysteme benötigen in der Regel bis zu 30 Sekunden, um eine Änderung zu erkennen.

Die zu erwartenden Übernahmezeiten, während denen das Storage-System keine Daten für eine Applikationsumgebung bereitstellen kann, sind in der folgenden Tabelle aufgeführt. Es sollte keine Fehler in einer Applikationsumgebung geben, das Takeover sollte stattdessen als kurze Pause bei der I/O-Verarbeitung erscheinen.

	NFS	AFF	ASA
Geplante Übernahme	15 Sek.	6-10 Sek.	2-3 Sek.
Ungeplante Übernahme	30 Sek.	6-10 Sek.	2-3 Sek.

Prüfsummen und Datenintegrität

ONTAP und die unterstützten Protokolle umfassen mehrere Funktionen zum Schutz der Integrität der Oracle-Datenbank, darunter sowohl Daten im Ruhezustand als auch Daten, die über das Netzwerk übertragen werden.

Die logische Datensicherung innerhalb von ONTAP setzt sich aus drei Kernanforderungen zusammen:

- Daten müssen vor Datenbeschädigung geschützt werden.
- Die Daten müssen vor Laufwerksausfällen geschützt werden.
- Änderungen an Daten müssen vor Verlust geschützt werden.

Diese drei Anforderungen werden in den folgenden Abschnitten erläutert.

Netzwerkkorruption: Prüfsummen

Die grundlegendste Stufe des Datenschutzes ist die Prüfsumme, die einen speziellen Fehler erkennenden Code ist, der neben den Daten gespeichert wird. Eine Beschädigung der Daten bei der Netzwerkübertragung wird mit Hilfe einer Prüfsumme und in einigen Fällen mehreren Prüfsummen erkannt.

Ein FC-Frame enthält beispielsweise eine Form der Prüfsumme, die als zyklische Redundanzprüfung (CRC, Cyclic Redundancy Check) bezeichnet wird, um sicherzustellen, dass die Nutzlast während der Übertragung nicht beschädigt ist. Der Sender sendet sowohl die Daten als auch den CRC der Daten. Der Empfänger eines FC-Frames berechnet den CRC der empfangenen Daten neu, um sicherzustellen, dass er mit dem übertragenen CRC übereinstimmt. Wenn der neu berechnete CRC nicht mit dem CRC übereinstimmt, der dem Frame zugeordnet ist, sind die Daten beschädigt und der FC-Frame wird verworfen oder abgelehnt. Eine iSCSI-I/O-Operation umfasst Prüfsummen auf TCP/IP- und Ethernet-Ebenen und kann für zusätzlichen Schutz optional auch den CRC-Schutz auf der SCSI-Schicht beinhalten. Jede Bit-Beschädigung auf dem Kabel wird von der TCP-Schicht oder IP-Schicht erkannt, was zu einer erneuten Übertragung des Pakets führt. Wie bei FC führen Fehler im SCSI CRC zu einem Verwerfen oder Zurückweisen des Vorgangs.

Laufwerkbeschädigungen: Prüfsummen

Mit Prüfsummen wird auch die Integrität der auf Laufwerken gespeicherten Daten überprüft. Auf Laufwerke geschriebene Datenblöcke werden mit einer Prüfsummenfunktion gespeichert, die eine unvorhersehbare Anzahl ergibt, die mit den Originaldaten verknüpft ist. Wenn Daten vom Laufwerk gelesen werden, wird die Prüfsumme neu berechnet und mit der gespeicherten Prüfsumme verglichen. Wenn sie nicht übereinstimmt, sind die Daten beschädigt und müssen von der RAID-Schicht wiederhergestellt werden.

Datenbeschädigung: Verlorene Schreibvorgänge

Eine der schwierigsten Arten von Korruption ist ein verlorenes oder falsch geschaltenes Schreiben. Wenn ein Schreibvorgang bestätigt wird, muss er an der richtigen Stelle auf das Medium geschrieben werden. Datenbeschädigungen lassen sich mithilfe einer einfachen Prüfsumme, die mit den Daten gespeichert wurde, relativ einfach erkennen. Wenn der Schreibvorgang jedoch einfach verloren geht, dann könnte die vorherige Version der Daten noch existieren und die Prüfsumme wäre korrekt. Wenn der Schreibvorgang an einem falschen physischen Speicherort platziert wird, ist die zugehörige Prüfsumme erneut für die gespeicherten Daten gültig, auch wenn der Schreibvorgang andere Daten zerstört hat.

Die Lösung für diese Herausforderung ist wie folgt:

- Ein Schreibvorgang muss Metadaten enthalten, die den Speicherort angeben, an dem der Schreibvorgang erwartungsgemäß gefunden werden soll.
- Ein Schreibvorgang muss eine Art Versionskennung enthalten.

Wenn ONTAP einen Block schreibt, schließt er Daten ein, zu denen der Block gehört. Wenn ein nachfolgender Lesezugriff einen Block identifiziert, der jedoch aufgrund der Metadaten zu Standort 123 gehört, als er an Position 456 gefunden wurde, wurde der Schreibvorgang fehlgestellt.

Es ist schwieriger, einen vollständig verlorenen Schreibvorgang zu erkennen. Die Erklärung ist sehr kompliziert, aber im Wesentlichen speichert ONTAP Metadaten so, dass ein Schreibvorgang zu Updates an zwei verschiedenen Orten auf den Laufwerken führt. Wenn ein Schreibvorgang verloren geht, werden bei einem nachfolgenden Lesen der Daten und der zugehörigen Metadaten zwei unterschiedliche Versionsidentitäten angezeigt. Dies zeigt an, dass der Schreibvorgang vom Laufwerk nicht abgeschlossen wurde.

Verloren gegangene und falsch verlegte Schreibvorgänge sind äußerst selten, doch steigt mit zunehmendem Laufwerksanzahl und steigenden Datenmengen der Datensätze das Risiko. Jedes Storage-System, das Datenbank-Workloads unterstützt, sollte die verlorener Schreibschutz enthalten.

Laufwerksausfälle: RAID, RAID DP und RAID-TEC

Wenn ein Datenblock auf einem Laufwerk erkannt wird, dass er beschädigt ist oder das gesamte Laufwerk ausfällt und nicht verfügbar ist, müssen die Daten wiederhergestellt werden. Dies wird in ONTAP mithilfe von Paritätslaufwerken durchgeführt. Die Daten werden auf mehreren Datenlaufwerken verteilt und anschließend Paritätsdaten generiert. Diese wird getrennt von den Originaldaten gespeichert.

ONTAP verwendete ursprünglich RAID 4, das für jede Gruppe von Datenlaufwerken ein Single-Parity-Laufwerk verwendet. Das Ergebnis war, dass ein Laufwerk in der Gruppe ausfallen konnte, ohne dass es zu Datenverlust kam. Bei einem Ausfall des Paritätslaufwerks wurden keine Daten beschädigt und ein neues Paritätslaufwerk erstellt. Wenn ein einzelnes Datenlaufwerk ausfällt, können die verbleibenden Laufwerke zusammen mit dem Paritätslaufwerk verwendet werden, um die fehlenden Daten neu zu generieren.

Bei geringen Laufwerksanzahl war die statistische Wahrscheinlichkeit, dass zwei Laufwerke gleichzeitig ausfallen, vernachlässigbar. Mit wachsenden Laufwerkskapazitäten hat sich auch die Zeit entwickelt, die für die Wiederherstellung von Daten nach einem Laufwerksausfall benötigt wird. Dadurch erhöht sich das Zeitfenster, in dem ein zweiter Laufwerksausfall zum Datenverlust führen würde. Darüber hinaus erzeugt der Neuerstellungsvorgang eine Menge zusätzlicher I/O auf den verbleibenden Laufwerken. Mit zunehmendem Festplattenalter steigt auch das Risiko, dass die zusätzliche Last zu einem zweiten Laufwerksausfall führt. Selbst wenn das Risiko eines Datenverlusts mit der fortgesetzten Nutzung von RAID 4 nicht Anstieg, würden die Folgen eines Datenverlusts schwerwiegender. Je mehr Daten im Falle eines Ausfalls einer RAID-Gruppe verloren gehen würden, desto länger würde die Wiederherstellung der Daten dauern, wodurch die Unterbrechung des Geschäftsbetriebs käme.

Aus diesen Problemen entwickelte NetApp die NetApp RAID DP-Technologie, eine Variante von RAID 6. Diese Lösung umfasst zwei Paritätslaufwerke, d. h., zwei beliebige Laufwerke einer RAID-Gruppe können ohne Datenverlust ausfallen. Die Größe der Laufwerke wurde weiter vergrößert, wodurch NetApp schließlich die NetApp RAID-TEC-Technologie entwickelt hat, wodurch ein drittes Paritätslaufwerk eingeführt wird.

Einige bewährte Verfahren für historische Datenbanken empfehlen die Verwendung von RAID-10, auch als Striped Mirroring bekannt. Dies bietet weniger Datensicherheit als RAID DP, da mehrere zwei-Festplatten-Fehlerszenarien auftreten, während es in RAID DP keine gibt.

Es gibt auch einige historische Best Practices für Datenbanken, die darauf hinweisen, dass RAID-10 aufgrund von Performance-Bedenken den Optionen RAID-4/5/6 vorzuziehen ist. Diese Empfehlungen beziehen sich manchmal auf einen RAID-Abzug. Obwohl diese Empfehlungen in der Regel richtig sind, gelten sie nicht für die Implementierungen von RAID innerhalb von ONTAP. Die Leistungsbedenken beziehen sich auf die Paritäts-Regeneration. Bei herkömmlichen RAID-Implementierungen müssen bei der Verarbeitung der routinemäßigen, zufälligen Schreibvorgänge durch eine Datenbank mehrere Lesezugriffe auf die Festplatte durchgeführt werden, um die Paritätsdaten neu zu generieren und den Schreibvorgang abzuschließen. Der Abzug wird definiert als die zusätzlichen Lese-IOPS, die zum Ausführen von Schreibvorgängen erforderlich sind.

Bei ONTAP kommt es nicht zu RAID-Einbußen, da Schreibvorgänge in den Speicher ausgelagert werden, wo Parität erzeugt wird und dann als einzelner RAID-Stripe auf die Festplatte geschrieben wird. Zum Abschließen des Schreibvorgangs sind keine Lesevorgänge erforderlich.

Zusammengefasst bieten RAID DP und RAID-TEC im Vergleich zu RAID 10 viel mehr nutzbare Kapazität, besseren Schutz vor Festplattenausfällen und keine Performance-Einbußen.

Schutz vor Hardware-Ausfällen: NVRAM

Jedes Storage-Array für Datenbank-Workloads muss Schreibvorgänge so schnell wie möglich durchführen. Darüber hinaus muss ein Schreibvorgang vor einem Verlust durch unerwartete Ereignisse, wie z. B. einen Stromausfall, geschützt werden. Das bedeutet, dass jeder Schreibvorgang sicher an mindestens zwei Orten gespeichert werden muss.

AFF und FAS Systeme vertrauen zur Erfüllung dieser Anforderungen auf NVRAM. Der Schreibvorgang funktioniert wie folgt:

1. Die eingehenden Schreibdaten werden im RAM gespeichert.
2. Die Änderungen, die an Daten auf Festplatte vorgenommen werden müssen, werden sowohl auf dem lokalen Node als auch auf dem Partner-Node in NVRAM eingetragen. NVRAM ist kein Schreib-Cache, sondern ein Journal, das einem Datenbank-Wiederherstellungsprotokoll ähnelt. Unter normalen Bedingungen wird sie nicht gelesen. Sie wird nur für die Wiederherstellung verwendet, z. B. nach einem Stromausfall während der I/O-Verarbeitung.
3. Der Schreibvorgang wird dann dem Host bestätigt.

Der Schreibvorgang in dieser Phase ist aus Sicht der Applikation abgeschlossen, und die Daten sind vor Verlust geschützt, da sie an zwei verschiedenen Standorten gespeichert werden. Schließlich werden die Änderungen auf die Festplatte geschrieben, doch dieser Prozess ist aus Sicht der Applikation bandextern, da er nach dem Quittieren des Schreibvorgangs auftritt und sich somit nicht auf die Latenz auswirkt. Dieser Prozess ist wieder ähnlich wie die Datenbankprotokollierung. Eine Änderung an der Datenbank wird so schnell wie möglich in den Wiederherstellungsprotokollen aufgezeichnet und die Änderung wird dann als festgeschrieben bestätigt. Die Updates der Datendateien erfolgen viel später und haben keinen direkten Einfluss auf die Geschwindigkeit der Verarbeitung.

Bei einem Controller-Ausfall übernimmt der Partner-Controller die erforderlichen Festplatten und gibt die

protokollierten Daten im NVRAM wieder, um I/O-Vorgänge, die beim Ausfall gerade ausgeführt wurden, wiederherzustellen.

Schutz vor Hardware-Ausfällen: 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 zum HA-Partner ausfällt, werden diese aktiven Daten nicht mehr gespiegelt.

Wenn der lokale NVRAM einen Fehler meldet, wird der Node heruntergefahren. Dieses Herunterfahren führt zu einem Failover auf einen HA-Partner-Controller. Es gehen keine Daten verloren, da der Controller den Schreibvorgang nicht bestätigt hat.

ONTAP lässt kein Failover zu, wenn die Daten nicht synchron sind, es sei denn, das Failover wird 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 erzwungen wird, da Datenbanken große interne Daten-Caches auf der Festplatte aufbewahren. Wenn ein erzwungenes Failover 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.

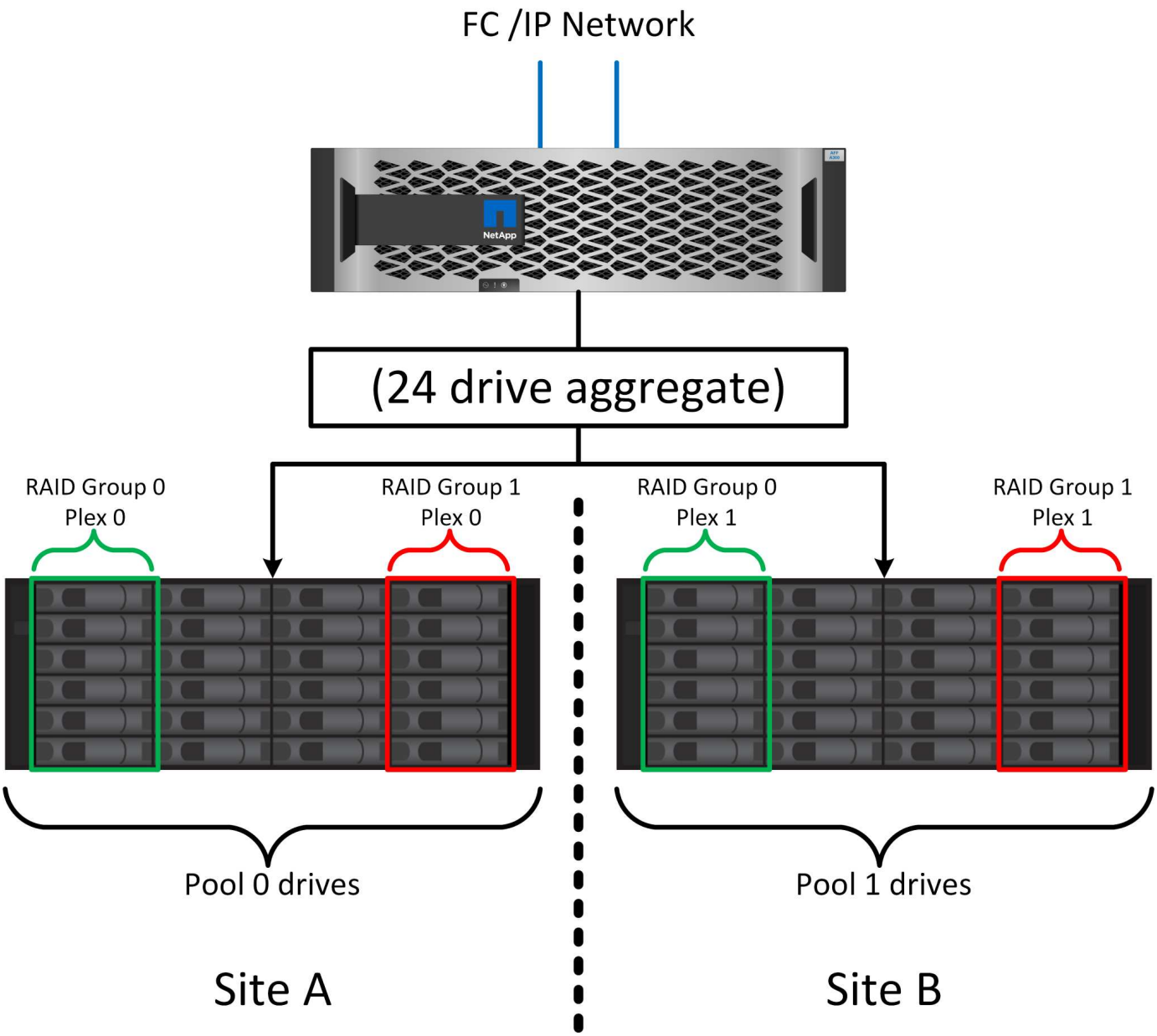
Um Daten aus dieser Situation zu schützen, können mit ONTAP Volumes für speziellen Schutz vor NVRAM-Ausfällen konfiguriert werden. Wenn dieser Schutzmechanismus ausgelöst wird, gelangt ein Volume in den Status „NVFAIL“. Dieser Status führt zu I/O-Fehlern, die dazu führen, dass Applikationen heruntergefahren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

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:

- 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
- 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
- Jeder Laufwerkssatz an einem bestimmten Standort wird automatisch als eine oder mehrere vollständig redundante RAID-DP- oder RAID-TEC-Gruppen konfiguriert, und zwar unabhängig vom Einsatz der Spiegelung. So wird eine kontinuierliche Datensicherung auch nach dem Verlust eines Standorts gewährleistet.



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 wird Ebenso enthält 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 getrennt, die von unabhängigen USV-Systemen mit Strom versorgt wurden.

Prüfsummen

Das Thema Prüfsummen ist von besonderem Interesse für DBAs, die es gewohnt sind, Oracle RMAN Streaming Backups zu Snapshot-basierten Backups zu verwenden. Eine Funktion von RMAN besteht darin, dass es während der Backups Integritätsprüfungen durchführt. Auch wenn dieses Feature einen gewissen Wert bietet, ist der Hauptvorteil für eine Datenbank, die nicht in einem modernen Storage-Array verwendet wird. Wenn physische Laufwerke für eine Oracle-Datenbank verwendet werden, ist es fast sicher, dass eine Beschädigung irgendwann auftritt, wenn die Laufwerke altern, ein Problem, das durch Array-basierte Prüfsummen in echten Storage-Arrays behoben wird.

Mit einem echten Storage-Array wird die Datenintegrität durch die Verwendung von Prüfsummen auf mehreren Ebenen gesichert. Wenn Daten in einem IP-basierten Netzwerk beschädigt sind, weist die TCP-Schicht (Transmission Control Protocol) die Paketdaten zurück und fordert eine erneute Übertragung an. Das FC-Protokoll umfasst Prüfsummen sowie eingekapselte SCSI-Daten. Nachdem es sich auf dem Array befindet, verfügt ONTAP über RAID- und Prüfsummenschutz. Es kann zu einer Beschädigung kommen, aber wie in den meisten Enterprise-Arrays wird sie erkannt und korrigiert. In der Regel fällt ein ganzes Laufwerk aus, was zu einer RAID-Neuerstellung führt, und die Datenbankintegrität bleibt davon unberührt. Es ist immer noch möglich, dass einzelne Bytes auf einem Laufwerk durch kosmische Strahlung oder fehlerhafte Blitzzellen beschädigt werden. In diesem Fall würde die Paritätsprüfung fehlschlagen, das Laufwerk würde ausfallen und eine RAID-Wiederherstellung würde beginnen. Auch hier bleibt die Datenintegrität erhalten. Die letzte Verteidigungslinie ist die Verwendung von Prüfsummen. Wenn zum Beispiel ein katastrophaler Firmware-Fehler auf einem Laufwerk Daten in einer Weise beschädigt, die irgendwie nicht durch eine RAID-Paritätsprüfung erkannt wurde, würde die Prüfsumme nicht übereinstimmen und ONTAP würde die Übertragung eines beschädigten Blocks verhindern, bevor die Oracle Datenbank den Block empfangen konnte.

Die Architektur der Oracle-Datendatei- und des Wiederherstellungsprotokolls wurde auch für höchste Datenintegrität entwickelt, selbst unter extremen Bedingungen. Auf der einfachsten Ebene enthalten Oracle-Blöcke Prüfsumme und grundlegende logische Prüfungen mit fast jedem I/O Wenn Oracle nicht abgestürzt ist oder einen Tablespace offline genommen hat, sind die Daten intakt. Der Grad der Datenintegritätsprüfung ist einstellbar und Oracle kann auch zur Bestätigung von Schreibvorgängen konfiguriert werden. Dadurch können fast alle Crash- und Ausfallszenarien wiederhergestellt werden. Im äußerst seltenen Fall einer nicht wiederherstellbaren Situation wird eine Beschädigung umgehend erkannt.

Die meisten NetApp-Kunden, die Oracle-Datenbanken einsetzen, beenden die Nutzung von RMAN und anderen Backup-Produkten nach der Migration zu Snapshot-basierten Backups. Es gibt nach wie vor Optionen, mit RMAN Recovery auf Blockebene mit SnapCenter durchgeführt werden kann. Allerdings werden RMAN, NetBackup und andere Produkte täglich nur gelegentlich verwendet, um monatliche oder vierteljährliche Archivkopien zu erstellen.

Einige Kunden wählen zu laufen `dbv` Regelmäßige Integritätsprüfungen der vorhandenen Datenbanken durchführen. NetApp rät von dieser Vorgehensweise ab, da dadurch unnötige I/O-Last erzeugt werden. Wie oben erwähnt, wenn die Datenbank zuvor keine Probleme hatte, die Chance von `dbv` Das Erkennen eines Problems ist nahezu gleich null, und dieses Dienstprogramm erzeugt eine sehr hohe sequenzielle I/O-Last auf dem Netzwerk und dem Speichersystem. Es sei denn, es gibt Grund zu der Annahme, dass Korruption vorhanden ist, wie die Offenlegung eines bekannten Oracle-Fehlers, gibt es keinen Grund, ausgeführt zu werden `dbv`.

Grundlagen von Backup und Recovery

Snapshot basierte Backups

Die Grundlage der Datensicherung für Oracle-Datenbanken auf ONTAP ist die NetApp Snapshot Technologie.

Die wichtigsten Werte sind:

- **Einfachheit.** Ein Snapshot ist eine schreibgeschützte Kopie des Inhalts eines Datencontainers zu einem bestimmten Zeitpunkt.
- **Effizienz.** Snapshots benötigen zum Zeitpunkt der Erstellung keinen Platz. Der Speicherplatz wird nur dann verbraucht, wenn Daten geändert werden.
- **Verwaltbarkeit.** Eine auf Snapshots basierende Backup-Strategie lässt sich einfach konfigurieren und verwalten, da Snapshots ein nativer Teil des Storage-Betriebssystems sind. Wenn das Speichersystem eingeschaltet ist, kann es Backups erstellen.
- **Skalierbarkeit.** bis zu 1024 Backups eines einzigen Dateicontainers und LUNs können beibehalten werden. Bei komplexen Datensätzen können diverse Daten-Container durch einen einzelnen, konsistenten Satz von Snapshots gesichert werden.
- Die Performance bleibt davon unberührt, ob ein Volume 1024 Snapshots enthält oder keine.

Viele Storage-Anbieter liefern zwar Snapshot-Technologie, doch ist die Snapshot Technologie bei ONTAP einzigartig und bietet in Enterprise-Applikations- und Datenbankumgebungen deutliche Vorteile:

- Snapshot Kopien sind Teil des zugrunde liegenden Write-Anywhere-Dateilayouts (WAFL). Es handelt sich nicht um ein Add-on oder eine externe Technologie. Dies vereinfacht das Management, da das Storage-System das Backup-System ist.
- Snapshot-Kopien beeinträchtigen die Performance nicht. Ausnahmen bilden Edge-Fälle, in denen so viele Daten in Snapshots gespeichert werden, dass sich das zugrunde liegende Storage-System füllt.
- Der Begriff „Konsistenzgruppe“ wird häufig verwendet, um eine Gruppierung von Storage-Objekten zu referenzieren, die als konsistente Sammlung von Daten gemanagt werden. Ein Snapshot eines bestimmten ONTAP Volumes stellt ein Konsistenzgruppenbackup dar.

ONTAP Snapshots lassen sich auch besser skalieren als bei Technologien von Mitbewerbern. Kunden können ohne Beeinträchtigung der Performance 5, 50 oder 500 Snapshots speichern. Derzeit sind in einem Volume maximal 1024 Snapshots zulässig. Wenn eine zusätzliche Snapshot-Aufbewahrung erforderlich ist, gibt es Optionen, die Snapshots an zusätzliche Volumes zu übergeben.

Daher ist die Sicherung eines auf ONTAP gehosteten Datensatzes einfach und hochskalierbar. Backups erfordern keine Verschiebung von Daten. Daher kann eine Backup-Strategie auf die Bedürfnisse des Unternehmens zugeschnitten werden und nicht auf die Beschränkungen der Netzwerkübertragungsraten, der großen Anzahl von Bandlaufwerken oder der Bereiche, in denen Festplatten bereitgestellt werden.

Ist ein Snapshot eine Sicherung?

Eine häufig gestellte Frage zur Verwendung von Snapshots als Datensicherungsstrategie ist die Tatsache, dass sich die „echten“ Daten und Snapshot-Daten auf denselben Laufwerken befinden. Der Verlust dieser Laufwerke würde sowohl zum Verlust der Primärdaten als auch des Backups führen.

Das ist ein berechtigte Anliegen. Lokale Snapshots werden für tägliche Backup- und Recovery-Anforderungen verwendet, in dieser Hinsicht ist der Snapshot ein Backup. Beinahe 99 % aller Recovery-Szenarien in NetApp Umgebungen basieren auf Snapshots, um selbst die anspruchsvollsten RTO-Anforderungen zu erfüllen.

Lokale Snapshots sollten jedoch nie die einzige Backup-Strategie sein. Deshalb bietet NetApp Technologien wie SnapMirror und SnapVault-Replizierung, um Snapshots schnell und effizient auf einen unabhängigen Laufwerkssatz zu replizieren. In einer richtig konzipierten Lösung mit Snapshots und Snapshot-Replikation kann die Verwendung von Tapes auf ein vierteljährliches Archiv minimiert oder ganz eliminiert werden.

Snapshot basierte Backups

Für die Sicherung Ihrer Daten gibt es viele Optionen für den Einsatz von ONTAP Snapshots. Snapshots bilden die Basis vieler anderer ONTAP Funktionen wie Replizierung, Disaster Recovery und Klonen. Eine vollständige Beschreibung der Snapshot-Technologie geht über den Umfang dieses Dokuments hinaus. Die folgenden Abschnitte bieten jedoch einen allgemeinen Überblick.

Es gibt zwei primäre Ansätze zum Erstellen eines Snapshots eines Datensatzes:

- Absturzkonsistente Backups
- Applikationskonsistente Backups

Ein absturzkonsistentes Backup eines Datensatzes bezieht sich auf die Erfassung der gesamten Datensatzstruktur zu einem bestimmten Zeitpunkt. Wenn der Datensatz in einem einzigen Volume gespeichert wird, ist der Vorgang einfach. Ein Snapshot kann jederzeit erstellt werden. Wenn ein Datensatz in mehreren Volumes gespeichert ist, muss ein Snapshot einer Konsistenzgruppe (CG) erstellt werden. Für das Erstellen von Snapshots von Konsistenzgruppen stehen verschiedene Optionen zur Verfügung, darunter NetApp SnapCenter-Software, native Funktionen von ONTAP-Konsistenzgruppen und vom Benutzer verwaltete Skripts.

Absturzkonsistente Backups kommen vor allem dann zum Einsatz, wenn die Recovery am Point-of-the-Backup ausreichend ist. Wenn ein granulareres Recovery erforderlich ist, sind in der Regel applikationskonsistente Backups erforderlich.

Das Wort „konsistent“ in „anwendungskonsistent“ ist oft eine Fehlbezeichnung. Das Platzieren einer Oracle-Datenbank in den Backup-Modus wird beispielsweise als applikationskonsistentes Backup bezeichnet, die Daten werden jedoch in keiner Weise konsistent oder stillgelegt. Die Daten ändern sich während des Backups weiterhin. Im Gegensatz dazu machen die meisten MySQL und Microsoft SQL Server Backups die Daten tatsächlich stillgelegt, bevor sie das Backup ausführen. VMware kann bestimmte Dateien konsistent machen oder auch nicht.

Konsistenzgruppen

Der Begriff „Konsistenzgruppe“ bezieht sich auf die Fähigkeit eines Speicherarrays, mehrere Speicherressourcen als ein einziges Image zu verwalten. Beispielsweise kann eine Datenbank aus 10 LUNs bestehen. Das Array muss in der Lage sein, diese 10 LUNs konsistent zu sichern, wiederherzustellen und zu replizieren. Eine Wiederherstellung ist nicht möglich, wenn die Images der LUNs zum Zeitpunkt des Backups nicht konsistent waren. Die Replikation dieser 10 LUNs erfordert, dass alle Replikate perfekt miteinander synchronisiert sind.

Der Begriff „Konsistenzgruppe“ wird nicht oft verwendet, wenn es um ONTAP geht, da Konsistenz immer eine Grundfunktion der Volume- und Aggregat-Architektur in ONTAP war. Viele andere Storage Arrays managen LUNs oder File-Systeme als einzelne Einheiten. Sie könnten aus Datenschutzgründen optional als „Konsistenzgruppe“ konfiguriert werden, dies ist jedoch ein zusätzlicher Schritt in der Konfiguration.

ONTAP war schon immer in der Lage, konsistente lokale und replizierte Images von Daten zu erfassen. Auch wenn die verschiedenen Volumes auf einem ONTAP-System normalerweise nicht formal als Konsistenzgruppe beschrieben werden, so sind sie doch das. Ein Snapshot dieses Volumes ist ein Konsistenzgruppenabbild, die Wiederherstellung dieses Snapshots ist eine Wiederherstellung der Konsistenzgruppe, und sowohl SnapMirror als auch SnapVault bieten Konsistenzgruppenreplikation.

Snapshots von Konsistenzgruppen

Konsistenzgruppen-Snapshots (cg-Snapshots) sind eine Erweiterung der grundlegenden ONTAP-Snapshot-Technologie. Bei einem standardmäßigen Snapshot-Vorgang wird ein konsistentes Image aller Daten innerhalb

eines einzelnen Volumes erstellt. In manchen Fällen ist es jedoch erforderlich, einen konsistenten Satz von Snapshots über mehrere Volumes und sogar über mehrere Storage-Systeme hinweg zu erstellen. Das Ergebnis ist ein Satz von Snapshots, die auf die gleiche Weise wie ein Snapshot von nur einem einzelnen Volume verwendet werden können. Sie können für die lokale Datenwiederherstellung verwendet, für Disaster Recovery-Zwecke repliziert oder als einheitliche konsistente Einheit geklont werden.

Die größte Verwendung von cg-Snapshots ist eine Datenbankumgebung mit einer Größe von ca. 1 PB und 12 Controllern. Die cg-Snapshots, die auf diesem System erstellt wurden, werden für Backups, Wiederherstellungen und Klonvorgänge verwendet.

Wenn ein Datensatz über mehrere Volumes verteilt und die Schreibreihenfolge beibehalten werden muss, wird meist automatisch ein cg-Snapshot von der ausgewählten Managementsoftware verwendet. Es besteht in solchen Fällen nicht die Notwendigkeit, die technischen Details von cg-Snapshots zu verstehen. Allerdings gibt es Situationen, in denen komplizierte Datensicherungsanforderungen eine detaillierte Kontrolle über den Datenschutz- und Replizierungsprozess erfordern. Einige Optionen sind Automatisierungs-Workflows oder der Einsatz benutzerdefinierter Skripte, um cg-Snapshot-APIs aufzurufen. Das Verständnis der besten Option und der Rolle von cg-Snapshot erfordert eine detailliertere Erläuterung der Technologie.

Die Erstellung eines Satzes von cg-Snapshots erfolgt in zwei Schritten:

1. Erstellung von Write Fencing auf allen Ziel-Volumes
2. Erstellen Sie Snapshots dieser Volumes im abgetrennten Zustand.

Schreibzaun wird seriell hergestellt. Das bedeutet, dass bei der Einrichtung des Fencing-Prozesses über mehrere Volumes hinweg die I/O-Schreibvorgänge auf dem ersten Volume in der Sequenz eingefroren werden, da sie weiterhin auf Volumes übertragen werden, die später angezeigt werden. Dies mag anfänglich möglicherweise gegen die Vorgabe verstoßen, die Schreibreihenfolge zu erhalten, gilt aber nur für I/O-Vorgänge, die asynchron auf dem Host ausgegeben werden und nicht von anderen Schreibvorgängen abhängen.

Beispielsweise kann eine Datenbank eine Vielzahl asynchroner Datendatei-Updates ausgeben und dem Betriebssystem ermöglichen, die I/O-Vorgänge neu zu ordnen und sie gemäß seiner eigenen Scheduler-Konfiguration abzuschließen. Die Reihenfolge dieser E/A-Typen kann nicht garantiert werden, da die Anwendung und das Betriebssystem bereits die Anforderung zur Wahrung der Schreibreihenfolge freigegeben haben.

Als Zählerbeispiel sind die meisten Datenbankprotokollierungsaktivitäten synchron. Die Datenbank fährt erst mit weiteren Protokollschreibvorgängen fort, nachdem der I/O-Vorgang bestätigt wurde und die Reihenfolge dieser Schreibvorgänge erhalten bleiben muss. Wenn ein Protokoll-I/O auf einem Volume mit Fencing ankommt, wird dies nicht bestätigt, und die Applikation blockiert weitere Schreibvorgänge. Ebenso ist der I/O der Filesystem-Metadaten in der Regel synchron. Beispielsweise darf ein Dateilösch nicht verloren gehen. Wenn ein Betriebssystem mit einem xfs-Dateisystem eine Datei und den I/O gelöscht hat, der die xfs-Dateisystemmetadaten aktualisiert hat, um den Verweis auf diese Datei zu entfernen, der auf einem umzäunten Volume gelandet ist, wird die Dateisystemaktivität angehalten. Dies garantiert die Integrität des Dateisystems während cg-Snapshot-Vorgängen.

Nach der Einrichtung von Write Fencing über die Ziel-Volumes hinweg sind sie für die Snapshot-Erstellung bereit. Die Snapshots müssen nicht genau zur gleichen Zeit erstellt werden, da der Zustand der Volumes aus einer abhängigen Schreibweise eingefroren wird. Um sich vor einem Fehler in der Anwendung zu schützen, die cg-Snapshots erstellt, enthält das anfängliche Write Fencing ein konfigurierbares Timeout, bei dem ONTAP die Fencing automatisch freigibt und die Schreibverarbeitung nach einer definierten Anzahl von Sekunden wieder aufnimmt. Wenn alle Snapshots erstellt werden, bevor die Zeitüberschreitung abgelaufen ist, dann ist der resultierende Snapshot-Satz eine gültige Konsistenzgruppe.

Abhängige Schreibreihenfolge

Aus technischer Sicht ist der Schlüssel zu einer Konsistenzgruppe die Aufrechterhaltung der Schreibreihenfolge und insbesondere der abhängigen Schreibreihenfolge. Beispielsweise wird eine Datenbank, die in 10 LUNs schreibt, gleichzeitig auf alle geschrieben. Viele Schreibvorgänge werden asynchron ausgegeben. Dies bedeutet, dass die Reihenfolge ihrer Fertigstellung unwichtig ist und die Reihenfolge ihrer Fertigstellung je nach Betriebssystem und Netzwerkverhalten variiert.

Einige Schreibvorgänge müssen auf der Festplatte vorhanden sein, bevor die Datenbank mit zusätzlichen Schreibvorgängen fortfahren kann. Diese kritischen Schreibvorgänge werden als abhängige Schreibvorgänge bezeichnet. Nachfolgende Schreib-I/O hängt davon ab, ob diese Schreibvorgänge auf der Festplatte vorhanden sind. Jeder Snapshot, jede Wiederherstellung oder Replikation dieser 10 LUNs muss sicherstellen, dass die abhängige Schreibreihenfolge gewährleistet ist. Dateisystemaktualisierungen sind ein weiteres Beispiel für Schreibvorgänge in Schreibreihenfolge. Die Reihenfolge, in der Dateisystemänderungen vorgenommen werden, muss beibehalten werden, oder das gesamte Dateisystem kann beschädigt werden.

Strategien

Es gibt zwei primäre Ansätze bei Snapshot-basierten Backups:

- Absturzkonsistente Backups
- Snapshot geschützte Hot-Backups

Ein absturzkonsistentes Backup einer Datenbank bezieht sich auf die Erfassung der gesamten Datenbankstruktur, einschließlich Datendateien, Wiederherstellungsprotokolle und Kontrolldateien zu einem bestimmten Zeitpunkt. Wenn die Datenbank in einem einzigen Volume gespeichert wird, ist der Vorgang einfach. Ein Snapshot kann jederzeit erstellt werden. Wenn eine Datenbank in mehreren Volumes gespeichert ist, muss ein Snapshot einer Konsistenzgruppe (CG) erstellt werden. Für das Erstellen von Snapshots von Konsistenzgruppen stehen verschiedene Optionen zur Verfügung, darunter NetApp SnapCenter-Software, native Funktionen von ONTAP-Konsistenzgruppen und vom Benutzer verwaltete Skripts.

Absturzkonsistente Snapshot Backups werden in erster Linie verwendet, wenn die Recovery eines bestimmten Backup ausreichend ist. Archivprotokolle können unter bestimmten Umständen eingesetzt werden. Wenn jedoch eine granularere zeitpunktgenaue Recovery erforderlich ist, ist ein Online-Backup vorzuziehen.

Das grundlegende Verfahren für ein Snapshot-basiertes Online-Backup ist wie folgt:

1. Platzieren Sie die Datenbank in `backup` Modus.
2. Erstellen Sie einen Snapshot aller Volumes, die Datendateien hosten.
3. Beenden `backup` Modus.
4. Führen Sie den Befehl aus `alter system archive log current` So erzwingen Sie die Protokollarchivierung.
5. Erstellen Sie Snapshots aller Volumes, die die Archivprotokolle hosten.

Dieses Verfahren ergibt einen Satz von Snapshots, die Datendateien im Backup-Modus enthalten, und die kritischen Archivprotokolle, die im Backup-Modus generiert wurden. Dies sind die beiden Anforderungen für das Recovery einer Datenbank. Dateien wie Kontrolldateien sollten ebenfalls aus Gründen der Bequemlichkeit geschützt werden, aber die einzige absolute Anforderung ist die Sicherung von Datendateien und Archivprotokollen.

Auch wenn unterschiedliche Kunden möglicherweise sehr unterschiedliche Strategien verfolgen, basieren fast alle diese Strategien letztendlich auf den unten erläuterten Prinzipien.

Snapshot-basierte Recovery

Beim Entwurf von Volume-Layouts für Oracle-Datenbanken ist die erste Entscheidung, ob die Volume-basierte VBSR-Technologie (NetApp SnapRestore) verwendet wird.

Mit Volume-basierten SnapRestore kann ein Volume fast sofort auf einen früheren Zeitpunkt zurückgesetzt werden. Da alle Daten auf dem Volume zurückgesetzt werden, ist VBSR möglicherweise nicht für alle Anwendungsfälle geeignet. Wenn beispielsweise eine gesamte Datenbank, einschließlich Datendateien, Wiederherstellungs- und Archivprotokolle, auf einem einzelnen Volume gespeichert ist und dieses Volume mit VBSR wiederhergestellt wird, gehen Daten verloren, da das neuere Archivprotokoll und die Wiederherstellungsdaten verworfen werden.

VBSR ist für die Wiederherstellung nicht erforderlich. Viele Datenbanken können mithilfe von dateibasiertem Single-File SnapRestore (SFSR) oder einfach durch Kopieren von Dateien aus dem Snapshot zurück in das aktive Dateisystem wiederhergestellt werden.

VBSR wird bevorzugt, wenn eine Datenbank sehr groß ist oder wenn sie so schnell wie möglich wiederhergestellt werden muss, und die Verwendung von VBSR erfordert die Isolierung der Datendateien. In einer NFS-Umgebung müssen die Datendateien einer bestimmten Datenbank in dedizierten Volumes gespeichert werden, die nicht durch andere Dateitypen kontaminiert sind. In einer SAN-Umgebung müssen Datendateien in dedizierten LUNs auf dedizierten Volumes gespeichert werden. Wenn ein Volume-Manager verwendet wird (einschließlich Oracle Automatic Storage Management [ASM]), muss die Festplattengruppe auch für Datendateien reserviert sein.

Werden Datendateien auf diese Weise isoliert, können sie in einen früheren Zustand zurückgesetzt werden, ohne andere Filesysteme zu beschädigen.

Snapshot Reserve

Für jedes Volume mit Oracle-Daten in einer SAN-Umgebung die `percent-snapshot-space` Sollte auf null gesetzt werden, da das Reservieren von Speicherplatz für einen Snapshot in einer LUN-Umgebung nicht nützlich ist. Wenn die fraktionale Reserve auf 100 eingestellt ist, benötigt ein Snapshot eines Volumes mit LUNs genug freien Platz im Volumen, ausgenommen die Snapshot-Reserve, um 100% Umsatz aller Daten aufzunehmen. Wenn die fraktionale Reserve auf einen niedrigeren Wert eingestellt ist, dann ist entsprechend weniger freier Speicherplatz erforderlich, schließt jedoch immer die Snapshot Reserve aus. Das bedeutet, dass der Speicherplatz der Snapshot-Reserve in einer LUN-Umgebung verschwendet wird.

In einer NFS-Umgebung gibt es zwei Optionen:

- Stellen Sie die ein `percent-snapshot-space` Basiert auf dem erwarteten Snapshot-Speicherplatzverbrauch.
- Stellen Sie die ein `percent-snapshot-space` Zur gemeinsamen Nutzung von Speicherplatz und Snapshots sowie zur Vermeidung und zum Management dieser Kapazitäten.

Mit der ersten Option `percent-snapshot-space` Wird auf einen Wert ungleich Null gesetzt, normalerweise etwa 20 %. Dieser Raum wird dann vor dem Benutzer ausgeblendet. Dieser Wert schafft jedoch keine Begrenzung der Auslastung. Wenn bei einer Datenbank mit einer Reservierung von 20 % 30 % anfällt, kann der Snapshot-Platz über die Grenze der 20-prozentigen Reserve hinauswachsen und nicht reservierten Speicherplatz belegen.

Der Hauptvorteil, wenn Sie eine Reserve auf einen Wert wie 20% setzen, besteht darin zu überprüfen, ob etwas Speicherplatz für Snapshots immer verfügbar ist. Bei einem 1-TB-Volume mit einer Reserve von 20 % wäre es beispielsweise nur einem Datenbankadministrator (DBA) möglich, 800 GB an Daten zu speichern. Diese Konfiguration garantiert mindestens 200 GB Speicherplatz für den Snapshot-Verbrauch.

Wenn `percent-snapshot-space` Ist auf null festgelegt, sodass der gesamte Speicherplatz im Volume für den Endbenutzer verfügbar ist, sodass bessere Sichtbarkeit gewährleistet wird. Ein DBA muss verstehen, dass ein 1-TB-Volume, das Snapshots nutzt, 1 TB Speicherplatz zwischen aktiven Daten und dem Snapshot-Umsatz gemeinsam genutzt wird.

Es gibt keine klare Präferenz zwischen Option 1 und Option 2 unter den Endbenutzern.

Snapshots von ONTAP und Drittanbietern

Oracle Doc ID 604683.1 erläutert die Anforderungen für die Snapshot-Unterstützung von Drittanbietern und die verschiedenen verfügbaren Optionen für Backup- und Wiederherstellungsvorgänge.

Der Drittanbieter muss sicherstellen, dass die Snapshots des Unternehmens den folgenden Anforderungen entsprechen:

- Snapshots müssen sich in die von Oracle empfohlenen Restore- und Recovery-Vorgänge integrieren.
- Snapshots müssen zum Zeitpunkt des Snapshots auch beim Absturz einer Datenbank konsistent sein.
- Die Schreibreihenfolge wird für jede Datei in einem Snapshot beibehalten.

Die Oracle Managementprodukte von ONTAP und NetApp erfüllen diese Anforderungen.

SnapRestore

Die schnelle Datenwiederherstellung in ONTAP anhand eines Snapshots wird durch die NetApp SnapRestore Technologie ermöglicht.

Wenn ein kritischer Datensatz nicht verfügbar ist, laufen die geschäftskritischen Prozesse ab. Tapes können beschädigt werden und selbst Restores aus festplattenbasierten Backups können die Übertragung über das Netzwerk verlangsamen. SnapRestore vermeidet diese Probleme durch eine nahezu sofortige Wiederherstellung der Datensätze. Selbst Datenbanken im Petabyte-Bereich lassen sich in wenigen Minuten vollständig wiederherstellen.

Es gibt zwei Arten von SnapRestore: Datei-/LUN-basiert und Volume-basiert.

- Einzelne Dateien oder LUNs lassen sich innerhalb von Sekunden wiederherstellen, egal ob es sich um eine 2-TB-LUN oder eine 4-KB-Datei handelt.
- Der Container von Dateien oder LUNs kann innerhalb von Sekunden wiederhergestellt werden, egal ob es sich um 10 GB oder 100 TB an Daten handelt.

Ein „Container mit Dateien oder LUNs“ würde sich normalerweise auf ein FlexVol Volume beziehen. Beispielsweise können Sie 10 LUNs aufweisen, aus denen sich eine LVM-Festplattengruppe in einem einzelnen Volume befindet. Alternativ kann ein Volume die NFS-Home-Verzeichnisse von 1000 Benutzern speichern. Anstatt für jede einzelne Datei oder jedes LUN einen Wiederherstellungsvorgang auszuführen, können Sie das gesamte Volume als einzelnen Vorgang wiederherstellen. Der Prozess funktioniert auch mit horizontal skalierbaren Containern, die mehrere Volumes enthalten, wie z. B. eine FlexGroup oder eine ONTAP-Konsistenzgruppe.

Der Grund, warum SnapRestore so schnell und effizient arbeitet, liegt in der Natur eines Snapshots, der im Wesentlichen eine parallele schreibgeschützte Ansicht der Inhalte eines Volumes zu einem bestimmten Zeitpunkt ist. Aktive Blöcke sind die realen Blöcke, die geändert werden können, während der Snapshot eine schreibgeschützte Ansicht des Status der Blöcke ist, die die Dateien und LUNs zum Zeitpunkt der Snapshot-Erstellung ausmachen.

ONTAP erlaubt nur schreibgeschützten Zugriff auf Snapshot-Daten, die Daten können jedoch mit SnapRestore reaktiviert werden. Der Snapshot wird als Lese-/Schreibansicht der Daten wieder aktiviert und gibt die Daten in ihren vorherigen Zustand zurück. SnapRestore kann auf Volume- oder Dateiebene betrieben werden. Die Technologie ist im Wesentlichen die gleiche mit ein paar geringfügigen Unterschiede im Verhalten.

Volume SnapRestore

Volume-basierte SnapRestore stellt das gesamte Datenvolumen in einen früheren Zustand zurück. Dieser Vorgang erfordert keine Datenverschiebung, d. h., der Wiederherstellungsprozess erfolgt im Wesentlichen unmittelbar, obwohl die Verarbeitung der API- oder CLI-Vorgänge einige Sekunden dauern kann. Die Wiederherstellung von 1 GB Daten ist nicht komplizierter und zeitaufwändiger als die Wiederherstellung von 1 PB Daten. Diese Funktion ist der Hauptgrund dafür, dass viele Enterprise-Kunden zu ONTAP Storage-Systemen migrieren. Die RTO wird in Sekunden für selbst größte Datensätze gemessen.

Ein Nachteil von Volume-basierten SnapRestore ist die Tatsache, dass Änderungen innerhalb eines Volumes im Laufe der Zeit kumuliert werden. Daher sind jeder Snapshot und die Daten der aktiven Datei von den bis zu diesem Zeitpunkt vorgenommenen Änderungen abhängig. Das Zurücksetzen eines Volumes in einen früheren Zustand bedeutet, dass alle nachfolgenden Änderungen, die an den Daten vorgenommen wurden, verworfen werden. Weniger offensichtlich ist jedoch, dass dies nachträglich erstellte Snapshots einschließt. Das ist nicht immer wünschenswert.

Beispielsweise kann in einem SLA für die Datenaufbewahrung eine nächtliche Sicherung von 30 Tagen festgelegt werden. Wenn ein Datensatz auf einen vor fünf Tagen mit Datenträger SnapRestore erstellten Snapshot wiederhergestellt wird, werden alle in den letzten fünf Tagen erstellten Snapshots verworfen und dies verstößt gegen den SLA.

Es gibt eine Reihe von Optionen, um diese Einschränkung zu beheben:

1. Daten können von einem früheren Snapshot kopiert werden, anstatt eine SnapRestore des gesamten Volumes durchzuführen. Diese Methode eignet sich am besten für kleinere Datensätze.
2. Ein Snapshot kann geklont und nicht wiederhergestellt werden. Die Einschränkung dieses Ansatzes besteht darin, dass der Quell-Snapshot eine Abhängigkeit des Klons ist. Daher kann sie nur gelöscht oder in ein unabhängiges Volume aufgesplittet werden.
3. Verwendung von dateibasiertem SnapRestore.

File SnapRestore

Bei File-basierten SnapRestore handelt es sich um einen granulareren Snapshot-basierten Wiederherstellungsprozess. Anstatt den Status eines gesamten Volume zurückzusetzen, wird der Status einer einzelnen Datei oder LUN zurückgesetzt. Es müssen keine Snapshots gelöscht werden. Durch diesen Vorgang wird auch keine Abhängigkeit von einem vorherigen Snapshot erzeugt. Die Datei oder LUN ist im aktiven Volume sofort verfügbar.

Bei einem SnapRestore Restore einer Datei oder eines LUN sind keine Datenverschiebungen erforderlich. Einige interne Metadaten-Updates sind jedoch erforderlich, um abzubilden, dass die zugrunde liegenden Blöcke in einer Datei oder einem LUN jetzt sowohl in einem Snapshot als auch in dem aktiven Volume vorhanden sind. Die Performance sollte sich nicht auswirken, doch bei diesem Prozess wird die Erstellung von Snapshots blockiert, bis dieser abgeschlossen ist. Die Verarbeitungsrate beträgt ca. 5 Gbit/s (18 TB/Stunde), basierend auf der Gesamtgröße der wiederhergestellten Dateien.

Online-Backups

Zwei Datensätze sind erforderlich, um eine Oracle Datenbank im Backup-Modus zu schützen und wiederherzustellen. Beachten Sie, dass dies nicht die einzige Oracle-

Backup-Option ist, aber es ist die häufigste.

- Ein Snapshot der Datendateien im Backup-Modus
- Die Archivprotokolle, die erstellt wurden, während sich die Datendateien im Backup-Modus befanden

Wenn eine vollständige Recovery einschließlich aller festgeschriebenen Transaktionen notwendig ist, ist ein dritter Artikel erforderlich:

- Ein Satz aktueller Wiederherstellungsprotokolle

Es gibt eine Reihe von Möglichkeiten, die Recovery eines Online-Backups zu fördern. Viele Kunden stellen Snapshots mithilfe der ONTAP CLI wieder her und verwenden dann Oracle RMAN oder sqlplus, um die Recovery abzuschließen. Dies ist besonders bei großen Produktionsumgebungen der Fall, in denen die Wahrscheinlichkeit und Häufigkeit der Wiederherstellung von Datenbanken äußerst gering ist und alle Wiederherstellungsverfahren von einem erfahrenen DBA durchgeführt werden. Für die vollständige Automatisierung verfügen Lösungen wie NetApp SnapCenter über ein Oracle Plug-in mit Befehlszeile und grafischer Benutzeroberfläche.

Einige große Kunden haben einen einfacheren Ansatz verfolgt, indem sie einfache Skripte auf den Hosts konfigurieren, um die Datenbanken zu einem bestimmten Zeitpunkt in den Backup-Modus zu versetzen, um einen geplanten Snapshot vorzubereiten. Planen Sie beispielsweise den Befehl `alter database begin backup` Um 23:58 `alter database end backup` Um 00:02 Uhr, und planen Sie dann Snapshots direkt auf dem Speichersystem um Mitternacht. Das Ergebnis ist eine einfache, hochgradig skalierbare Backup-Strategie, für die keine externe Software oder Lizenzen erforderlich sind.

Datenlayout

Am einfachsten ist es, Datendateien in einem oder mehreren dedizierten Volumes zu isolieren. Sie müssen durch einen anderen Dateityp nicht kontaminiert sein. Dadurch soll sichergestellt werden, dass die Datendatei-Volumes über einen SnapRestore-Vorgang schnell wiederhergestellt werden können, ohne dass ein wichtiges Wiederherstellungsprotokoll, eine Steuerdatei oder ein Archivprotokoll zerstört werden.

SAN hat ähnliche Anforderungen für die Isolation von Datendateien in dedizierten Volumes. Bei einem Betriebssystem wie Microsoft Windows kann ein einzelnes Volume mehrere Datendatei-LUNs mit jeweils einem NTFS-Filesystem enthalten. Bei anderen Betriebssystemen gibt es in der Regel einen logischen Volume Manager. Mit Oracle ASM wäre es beispielsweise am einfachsten, die LUNs einer ASM-Laufwerksgruppe auf ein einzelnes Volume zu beschränken, das als Einheit gesichert und wiederhergestellt werden kann. Wenn aus Gründen der Performance oder des Kapazitätsmanagements zusätzliche Volumes erforderlich sind, vereinfacht sich das Management durch die Erstellung einer zusätzlichen Festplattengruppe auf dem neuen Volume.

Wenn diese Richtlinien befolgt werden, können Snapshots direkt auf dem Speichersystem geplant werden, ohne dass ein Snapshot einer Konsistenzgruppe erforderlich ist. Der Grund hierfür liegt darin, dass für Oracle-Backups keine Datendateien gleichzeitig gesichert werden müssen. Das Online-Backup-Verfahren wurde entwickelt, damit Datendateien weiterhin aktualisiert werden können, da sie im Laufe der Stunden langsam auf Tape gestreamt werden.

Eine Komplikation entsteht in Situationen wie der Verwendung einer ASM-Datenträgergruppe, die auf Volumes verteilt ist. In diesen Fällen muss ein cg-Snapshot ausgeführt werden, um sicherzustellen, dass die ASM-Metadaten über alle zusammengehörigen Volumes hinweg konsistent sind.

Achtung: Überprüfen Sie, dass der ASM `spfile` Und `passwd` Die Dateien befinden sich nicht in der Festplattengruppe, in der die Datendateien gehostet werden. Dies beeinträchtigt die Fähigkeit, Datendateien und nur Datendateien selektiv wiederherzustellen.

Verfahren zur lokalen Wiederherstellung – NFS

Dieses Verfahren kann manuell oder über eine Anwendung wie SnapCenter gesteuert werden. Das Grundverfahren ist wie folgt:

1. Fahren Sie die Datenbank herunter.
2. Stellen Sie die Datendatei-Volumes unmittelbar vor dem gewünschten Wiederherstellungspunkt auf den Snapshot wieder her.
3. Geben Sie Archivprotokolle bis zum gewünschten Punkt wieder.
4. Wiederholen Sie die aktuellen Wiederherstellungsprotokolle, wenn eine vollständige Wiederherstellung gewünscht wird.

Bei diesem Verfahren wird davon ausgegangen, dass die gewünschten Archivprotokolle noch im aktiven Dateisystem vorhanden sind. Ist dies nicht der Fall, müssen die Archivprotokolle wiederhergestellt werden oder rman/sqlplus kann zu den Daten im Snapshot-Verzeichnis geleitet werden.

Außerdem können Datendateien bei kleineren Datenbanken von einem Endbenutzer direkt aus wiederhergestellt werden `.snapshot` Verzeichnis ohne die Unterstützung von Automatisierungs-Tools oder Storage-Administratoren, ein auszuführen `snaprestore` Befehl.

Verfahren zur lokalen Wiederherstellung – SAN

Dieses Verfahren kann manuell oder über eine Anwendung wie SnapCenter gesteuert werden. Das Grundverfahren ist wie folgt:

1. Fahren Sie die Datenbank herunter.
2. Legen Sie die Festplattengruppe(n), die die Datendateien hosten, still. Die Vorgehensweise hängt vom gewählten Logical Volume Manager ab. Bei ASM muss die Datenträgergruppe demontieren. Bei Linux müssen die Dateisysteme demontiert und die logischen Volumes und Volume-Gruppen deaktiviert werden. Ziel ist es, alle Aktualisierungen auf der Zieldatentengruppe zu stoppen, die wiederhergestellt werden sollen.
3. Stellen Sie die Datendatei-Datenträgergruppen auf dem Snapshot unmittelbar vor dem gewünschten Wiederherstellungspunkt wieder her.
4. Reaktivieren Sie die neu wiederhergestellten Datenträgergruppen.
5. Geben Sie Archivprotokolle bis zum gewünschten Punkt wieder.
6. Wiederholen Sie alle Wiederherstellungsprotokolle, wenn eine vollständige Wiederherstellung gewünscht wird.

Bei diesem Verfahren wird davon ausgegangen, dass die gewünschten Archivprotokolle noch im aktiven Dateisystem vorhanden sind. Wenn dies nicht der Fall ist, müssen die Archivprotokolle wiederhergestellt werden, indem die Archivprotokoll-LUNs offline geschaltet und eine Wiederherstellung durchgeführt wird. Dies ist ebenfalls ein Beispiel, bei dem sich Archivprotokolle in dedizierte Volumes aufteilen lassen. Wenn die Archivprotokolle eine Volume-Gruppe mit Wiederherstellungsprotokollen gemeinsam nutzen, müssen die Wiederherstellungsprotokolle vor der Wiederherstellung des gesamten LUN-Satzes an eine andere Stelle kopiert werden. Dieser Schritt verhindert den Verlust dieser letzten aufgezeichneten Transaktionen.

Storage Snapshot optimierte Backups

Snapshot-basiertes Backup und Recovery wurden vor der Veröffentlichung von Oracle 12c noch einfacher, da eine Datenbank nicht im Hot-Backup-Modus platziert werden muss. Daraus ergibt sich die Möglichkeit, Snapshot basierte Backups direkt auf einem

Storage-System zu planen und dennoch eine vollständige oder zeitpunktgenaue Recovery durchzuführen.

Obwohl DBAs mit der Hot-Backup-Wiederherstellung vertrauter sind, ist es seit langem möglich, Snapshots zu verwenden, die nicht erstellt wurden, während sich die Datenbank im Hot-Backup-Modus befand. Für Oracle 10g und 11g waren während der Recovery zusätzliche manuelle Schritte erforderlich, um die Datenbankkonsistenz zu gewährleisten. Mit Oracle 12c, `sqlplus` Und `rman` Enthalten die zusätzliche Logik zur Wiedergabe von Archivprotokollen für Datendatei-Backups, die sich nicht im Hot-Backup-Modus befanden.

Wie bereits erwähnt, erfordert die Wiederherstellung eines Snapshot-basierten Hot-Backups zwei Datensätze:

- Ein Snapshot der Datendateien, der im Backup-Modus erstellt wurde
- Die Archivprotokolle, die generiert wurden, während sich die Datendateien im Hot-Backup-Modus befanden

Während der Recovery liest die Datenbank Metadaten aus den Datendateien, um die erforderlichen Archivprotokolle für die Recovery auszuwählen.

Storage Snapshot optimierte Recovery erfordert geringfügig unterschiedliche Datensätze, um die gleichen Ergebnisse zu erzielen:

- Ein Snapshot der Datendateien und eine Methode zur Identifizierung des Zeits, zu dem der Snapshot erstellt wurde
- Archivieren Sie Protokolle vom Zeitpunkt des letzten Datendatei-Kontrollpunkts bis zum genauen Zeitpunkt des Snapshots

Während der Recovery liest die Datenbank Metadaten aus den Datendateien, um das früheste erforderliche Archivprotokoll zu identifizieren. Eine vollständige oder zeitpunktgenaue Recovery kann durchgeführt werden. Bei einer zeitpunktgenauen Recovery ist es wichtig, die Zeit des Snapshots der Datendateien zu kennen. Der angegebene Wiederherstellungspunkt muss nach der Erstellungszeit der Snapshots liegen. NetApp empfiehlt, die Snapshot-Zeit um mindestens einige Minuten zu erweitern, um Uhrschwankungen zu berücksichtigen.

Ausführliche Informationen finden Sie in der Oracle-Dokumentation zum Thema „Recovery Using Storage Snapshot Optimization“, die in verschiedenen Versionen der Oracle 12c-Dokumentation verfügbar ist. Weitere Informationen zur Snapshot-Unterstützung von Drittanbietern finden Sie unter Oracle Document ID Doc ID 604683.1.

Datenlayout

Am einfachsten ist es, die Datendateien in einem oder mehreren dedizierten Volumes zu isolieren. Sie müssen durch einen anderen Dateityp nicht kontaminiert sein. Dadurch soll sichergestellt werden, dass die Datendatei-Volumes mit einem SnapRestore-Vorgang schnell wiederhergestellt werden können, ohne dass ein wichtiges Wiederherstellungsprotokoll, eine Steuerdatei oder ein Archivprotokoll zerstört werden.

SAN hat ähnliche Anforderungen für die Isolation von Datendateien in dedizierten Volumes. Bei einem Betriebssystem wie Microsoft Windows kann ein einzelnes Volume mehrere Datendatei-LUNs mit jeweils einem NTFS-Filesystem enthalten. Bei anderen Betriebssystemen gibt es in der Regel auch einen logischen Volume Manager. Mit Oracle ASM wäre es beispielsweise am einfachsten, Laufwerksgruppen auf ein einzelnes Volume zu beschränken, das als Einheit gesichert und wiederhergestellt werden kann. Wenn aus Gründen der Performance oder des Kapazitätsmanagements zusätzliche Volumes erforderlich sind, erleichtert die Erstellung einer zusätzlichen Laufwerksgruppe auf dem neuen Volume das Management.

Wenn diese Richtlinien befolgt werden, können Snapshots direkt auf ONTAP geplant werden, ohne dass ein Snapshot einer Konsistenzgruppe erforderlich ist. Der Grund hierfür liegt darin, dass Snapshot-optimierte

Backups keine gleichzeitige Sicherung von Datendateien erfordern.

Eine Komplikation entsteht in Situationen wie einer ASM-Datenträgergruppe, die auf Volumes verteilt ist. In diesen Fällen muss ein cg-Snapshot ausgeführt werden, um sicherzustellen, dass die ASM-Metadaten über alle zusammengehörigen Volumes hinweg konsistent sind.

[Hinweis]Vergewissern Sie sich, dass sich die ASM-Spfile- und Passwd-Dateien nicht in der Festplattengruppe befinden, die die Datendateien hostet. Dies beeinträchtigt die Fähigkeit, Datendateien und nur Datendateien selektiv wiederherzustellen.

Verfahren zur lokalen Wiederherstellung – NFS

Dieses Verfahren kann manuell oder über eine Anwendung wie SnapCenter gesteuert werden. Das Grundverfahren ist wie folgt:

1. Fahren Sie die Datenbank herunter.
2. Stellen Sie die Datendatei-Volumes unmittelbar vor dem gewünschten Wiederherstellungspunkt auf den Snapshot wieder her.
3. Geben Sie Archivprotokolle bis zum gewünschten Punkt wieder.

Bei diesem Verfahren wird davon ausgegangen, dass die gewünschten Archivprotokolle noch im aktiven Dateisystem vorhanden sind. Wenn dies nicht der Fall ist, müssen die Archivprotokolle wiederhergestellt werden, oder `rman` Oder `sqlplus` Kann auf die Daten im weitergeleitet werden `.snapshot` Verzeichnis.

Außerdem können Datendateien bei kleineren Datenbanken von einem Endbenutzer direkt aus wiederhergestellt werden `.snapshot` Directory ohne Unterstützung durch Automatisierungs-Tools oder einen Storage-Administrator, um einen SnapRestore-Befehl auszuführen.

Verfahren zur lokalen Wiederherstellung – SAN

Dieses Verfahren kann manuell oder über eine Anwendung wie SnapCenter gesteuert werden. Das Grundverfahren ist wie folgt:

1. Fahren Sie die Datenbank herunter.
2. Legen Sie die Festplattengruppe(n), die die Datendateien hosten, still. Die Vorgehensweise hängt vom gewählten Logical Volume Manager ab. Bei ASM muss die Datenträgergruppe demontieren. Bei Linux müssen die Dateisysteme getrennt und die logischen Volumes und Volume-Gruppen deaktiviert werden. Ziel ist es, alle Aktualisierungen auf der Zieldatentengruppe zu stoppen, die wiederhergestellt werden sollen.
3. Stellen Sie die Datendatei-Datenträgergruppen auf dem Snapshot unmittelbar vor dem gewünschten Wiederherstellungspunkt wieder her.
4. Reaktivieren Sie die neu wiederhergestellten Datenträgergruppen.
5. Geben Sie Archivprotokolle bis zum gewünschten Punkt wieder.

Bei diesem Verfahren wird davon ausgegangen, dass die gewünschten Archivprotokolle noch im aktiven Dateisystem vorhanden sind. Wenn dies nicht der Fall ist, müssen die Archivprotokolle wiederhergestellt werden, indem die Archivprotokoll-LUNs offline geschaltet und eine Wiederherstellung durchgeführt wird. Dies ist ebenfalls ein Beispiel, bei dem sich Archivprotokolle in dedizierte Volumes aufteilen lassen. Wenn die Archivprotokolle eine Volume-Gruppe mit Wiederherstellungsprotokollen gemeinsam nutzen, müssen die Wiederherstellungsprotokolle vor der Wiederherstellung des gesamten LUN-Satzes an eine andere Stelle kopiert werden, damit die letzten aufgezeichneten Transaktionen nicht verloren gehen.

Beispiel für eine vollständige Wiederherstellung

Angenommen, die Datendateien wurden beschädigt oder zerstört, und eine vollständige Recovery ist erforderlich. Das Verfahren ist wie folgt:

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover automatic;
Media recovery complete.
SQL> alter database open;
Database altered.
SQL>
```

Beispiel für eine zeitpunktgenaue Recovery

Der gesamte Wiederherstellungsvorgang erfolgt über einen einzigen Befehl: `recover automatic`.

Wenn eine Point-in-Time-Recovery erforderlich ist, muss der Zeitstempel der Snapshots bekannt sein und kann wie folgt identifiziert werden:

```
Cluster01::> snapshot show -vserver vserver1 -volume NTAP_oradata -fields
create-time
vserver   volume        snapshot   create-time
--------  ------------  ---------  ------------------------
vserver1  NTAP_oradata  my-backup  Thu Mar 09 10:10:06 2017
```

Die Erstellungszeit für Snapshots wird als 9. März und 10:10:06 aufgeführt. Um sicher zu sein, wird der Snapshot-Zeit eine Minute hinzugefügt:

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover database until time '09-MAR-2017 10:44:15' snapshot time '09-
MAR-2017 10:11:00';
```

Die Wiederherstellung ist nun gestartet. Es gab eine Snapshot-Zeit von 10:11:00, eine Minute nach der aufgezeichneten Zeit, um mögliche Taktabweichungen zu berücksichtigen, und eine Ziel-Recovery-Zeit von 10:44 an. Als Nächstes fordert sqlplus die Archivprotokolle an, die benötigt werden, um die gewünschte Wiederherstellungszeit von 10:44 zu erreichen.

```
ORA-00279: change 551760 generated at 03/09/2017 05:06:07 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_31_930813377.dbf
ORA-00280: change 551760 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 552566 generated at 03/09/2017 05:08:09 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_32_930813377.dbf
ORA-00280: change 552566 for thread 1 is in sequence #32
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 553045 generated at 03/09/2017 05:10:12 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_33_930813377.dbf
ORA-00280: change 553045 for thread 1 is in sequence #33
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 753229 generated at 03/09/2017 05:15:58 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_34_930813377.dbf
ORA-00280: change 753229 for thread 1 is in sequence #34
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
Log applied.
Media recovery complete.
SQL> alter database open resetlogs;
Database altered.
SQL>
```

Führen Sie die Wiederherstellung einer Datenbank mithilfe von Snapshots mit dem durch `recover automatic` Für Befehl ist keine spezifische Lizenzierung erforderlich, aber die zeitpunktgenaue Recovery mit `snapshot time` Erfordert die Oracle Advanced Compression-Lizenz.

Tools für Datenbankmanagement und Automatisierung

Der Hauptnutzen von ONTAP in einer Oracle Datenbankumgebung ergibt sich aus den zentralen ONTAP Technologien wie sofortige Snapshot Kopien, einfache SnapMirror Replizierung und die effiziente Erstellung von FlexClone Volumes.

In manchen Fällen erfüllt eine einfache Konfiguration dieser Kernfunktionen direkt in ONTAP die Anforderungen, für kompliziertere Anforderungen ist jedoch eine Orchestrierungsschicht erforderlich.

SnapCenter

SnapCenter ist das Vorzeigeprodukt für die Datensicherung von NetApp. Sie ähnelt im Hinblick auf die Durchführung von Datenbank-Backups den SnapManager Produkten. Sie wurde jedoch von Grund auf entwickelt, um bei NetApp Storage-Systemen eine zentrale Konsole für das Management der Daten zu bieten.

SnapCenter umfasst Grundfunktionen wie Snapshot-basierte Backups und Restores, SnapMirror und SnapVault Replizierung sowie weitere Funktionen, die für den skalierten Betrieb von Großunternehmen erforderlich sind. Zu diesen erweiterten Funktionen gehören eine erweiterte Funktion zur rollenbasierten Zugriffssteuerung (RBAC), RESTful APIs zur Integration in Orchestrierungsprodukte von Drittanbietern, unterbrechungsfreies, zentrales Management von SnapCenter Plug-ins auf Datenbank-Hosts und eine Benutzeroberfläche für Cloud-skalierbare Umgebungen.

RUHE

ONTAP enthält außerdem einen umfangreichen RESTful API-Satz. Drittanbieter sind so in der Lage, Datensicherungs- und Management-Applikationen mit enger Integration in ONTAP zu erstellen. Darüber hinaus kann die RESTful API von Kunden genutzt werden, die ihre eigenen Automatisierungs-Workflows und Dienstprogramme erstellen möchten.

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.

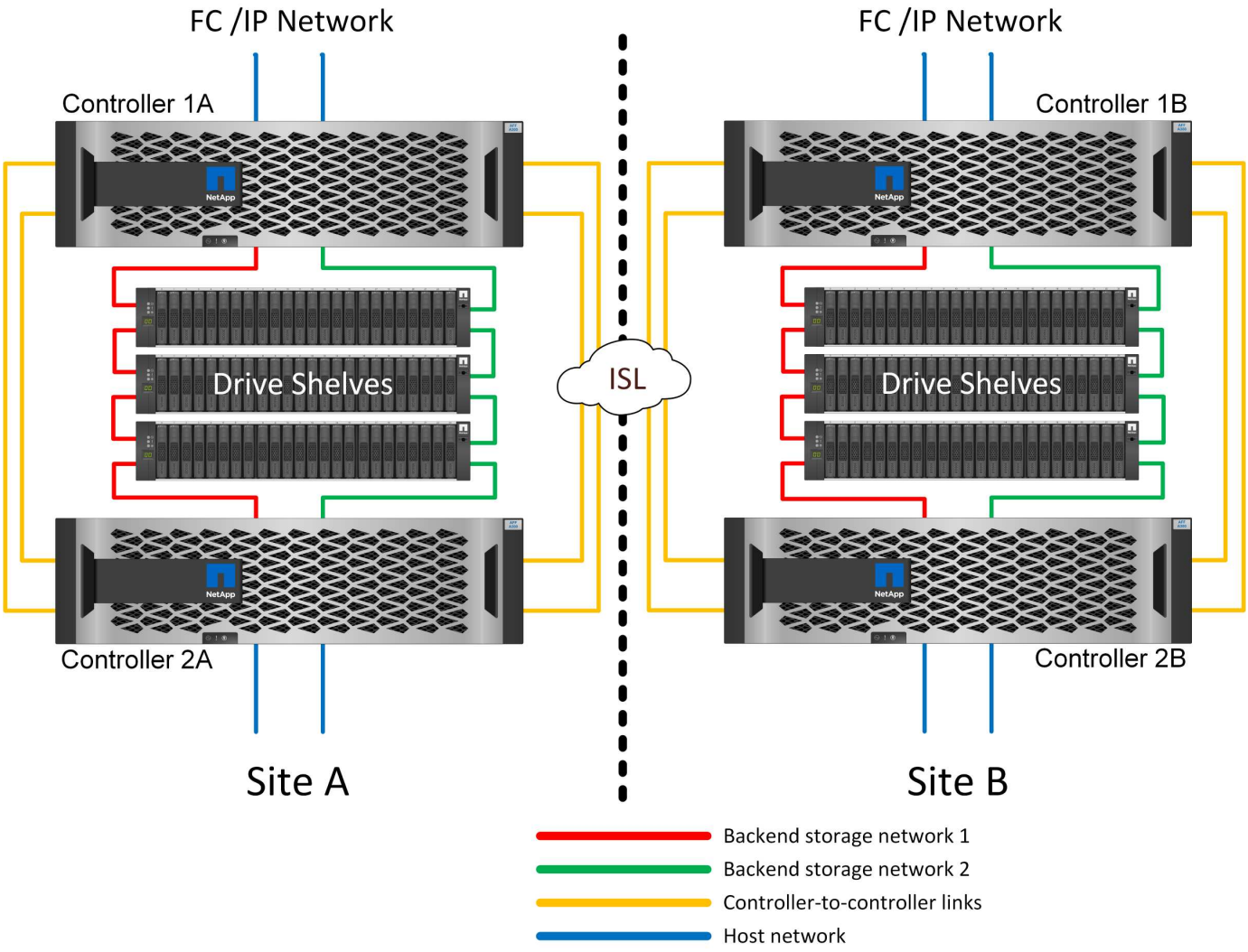
MetroCluster IP

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.

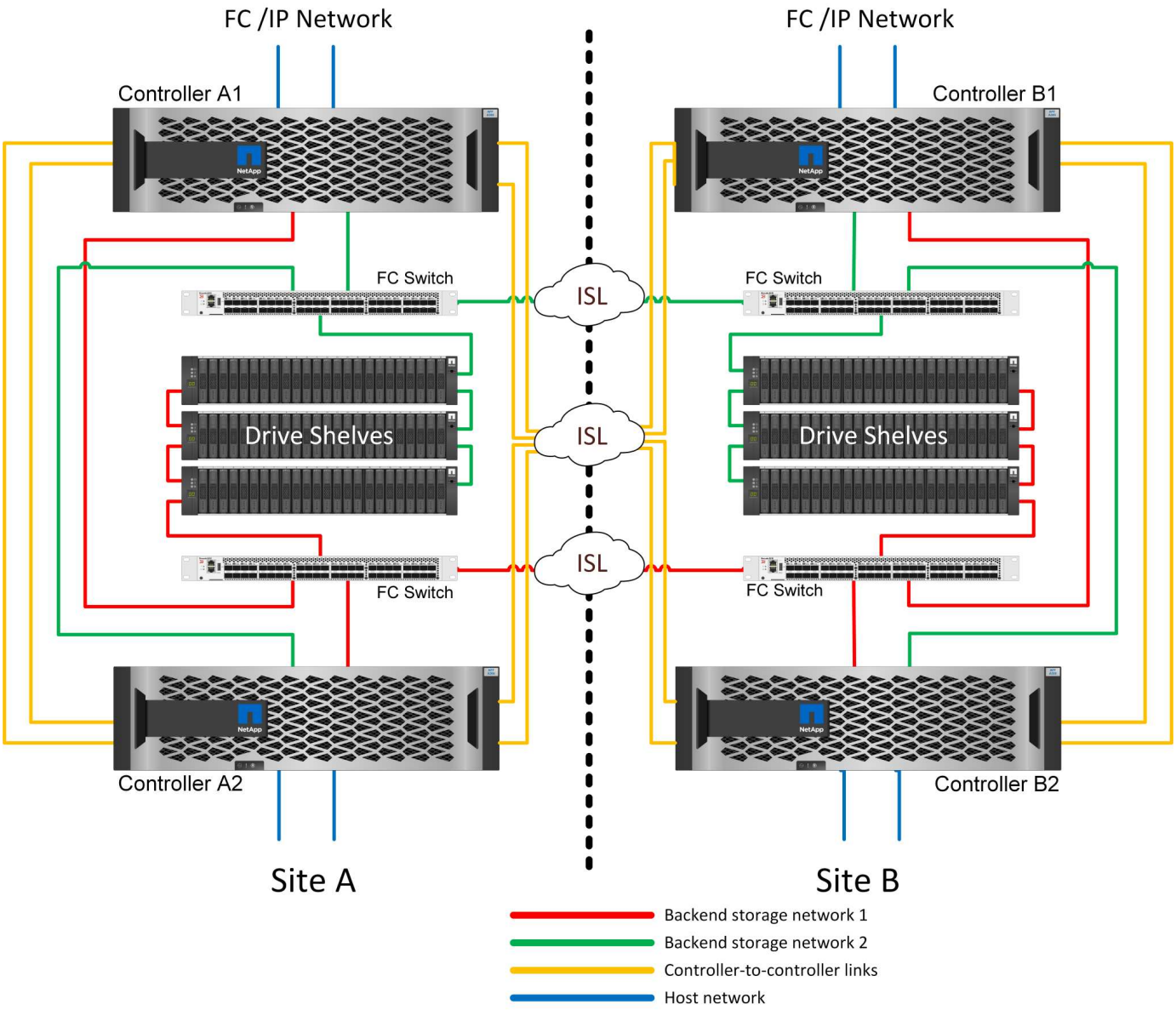
Auch die Architektur ist einfacher, da die einzigen standortübergreifenden Verbindungen für die Controller gelten. Bei FC SAN Attached MetroCluster schreibt ein Controller direkt auf die Laufwerke am entgegengesetzten Standort und benötigt somit zusätzliche SAN-Verbindungen, Switches und Bridges. Ein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



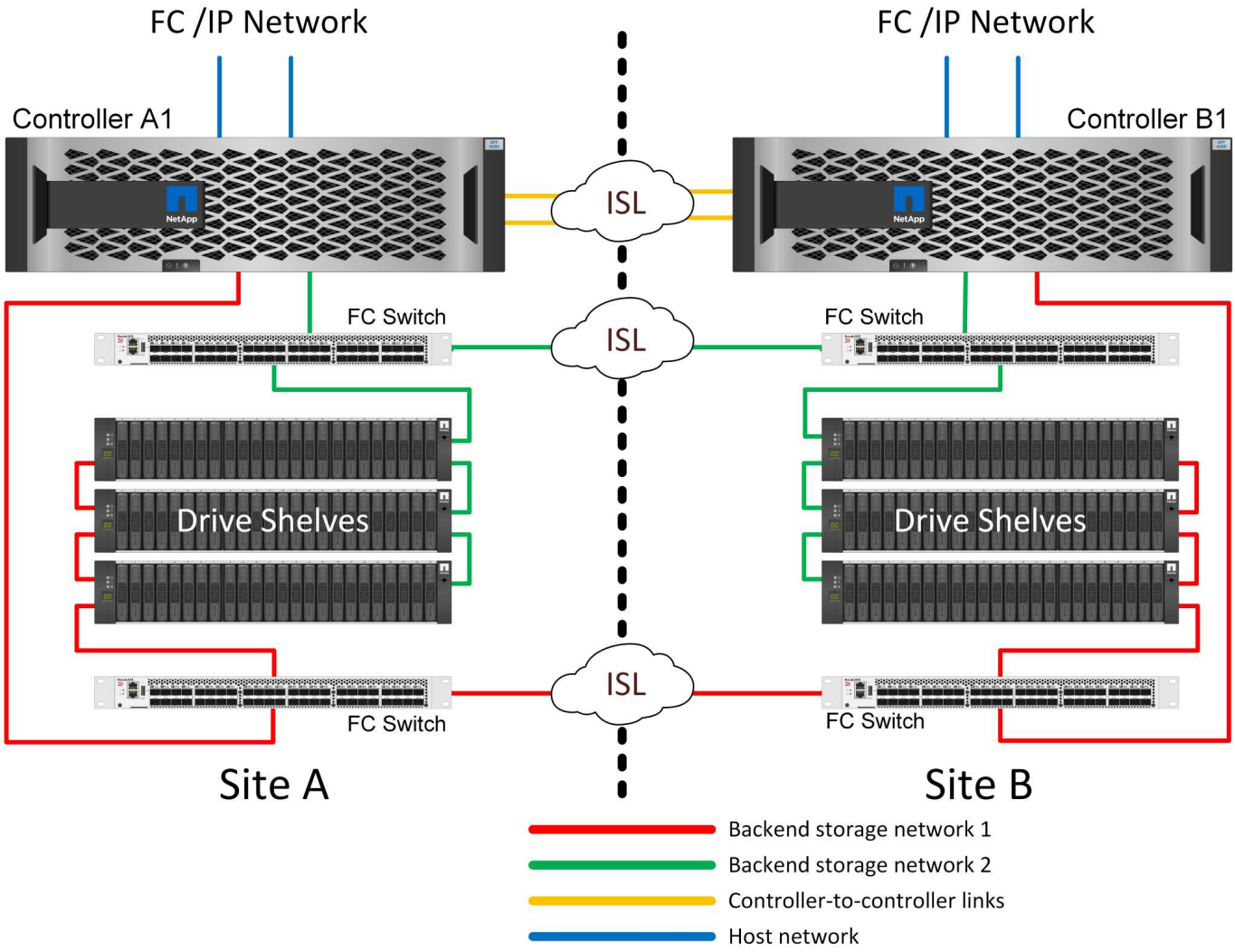
Einige Infrastrukturen mehrerer Standorte sind nicht für den aktiv/aktiv-Betrieb konzipiert, sondern werden eher als primärer Standort und Disaster-Recovery-Standort genutzt. In dieser Situation ist eine MetroCluster-Option für HA-Paare aus den folgenden Gründen im Allgemeinen vorzuziehen:

- 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

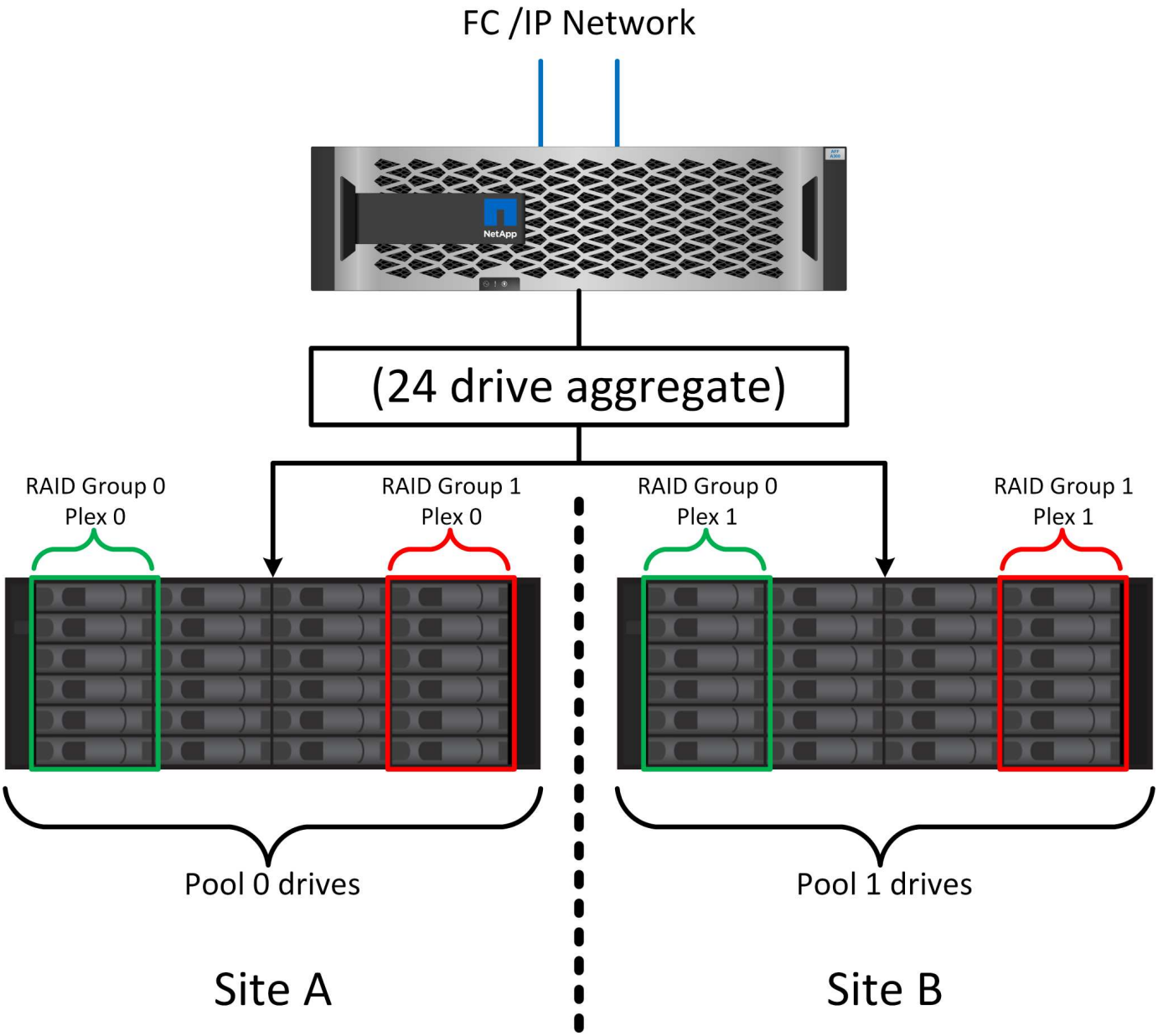
MetroCluster erweitert die NVRAM-Datensicherung auf folgende Weise:

- 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

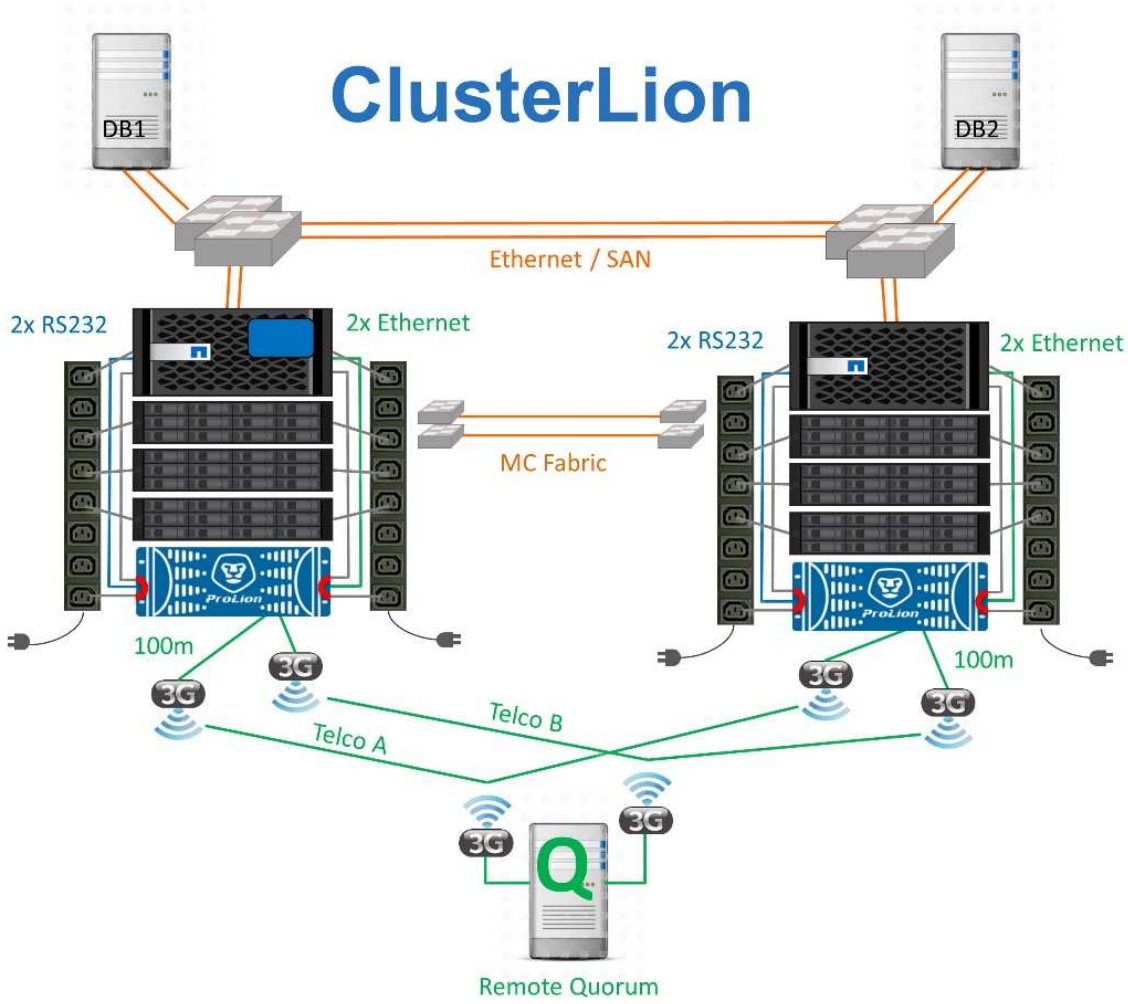
Der ONTAP Mediator wird mit MetroCluster IP und bestimmten anderen ONTAP-Lösungen verwendet. Es fungiert als herkömmlicher Tiebreaker Service, ähnlich wie die oben beschriebene MetroCluster Tiebreaker Software, verfügt aber auch über eine wichtige Funktion zum automatisierten, unbeaufsichtigten Switchover.

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.

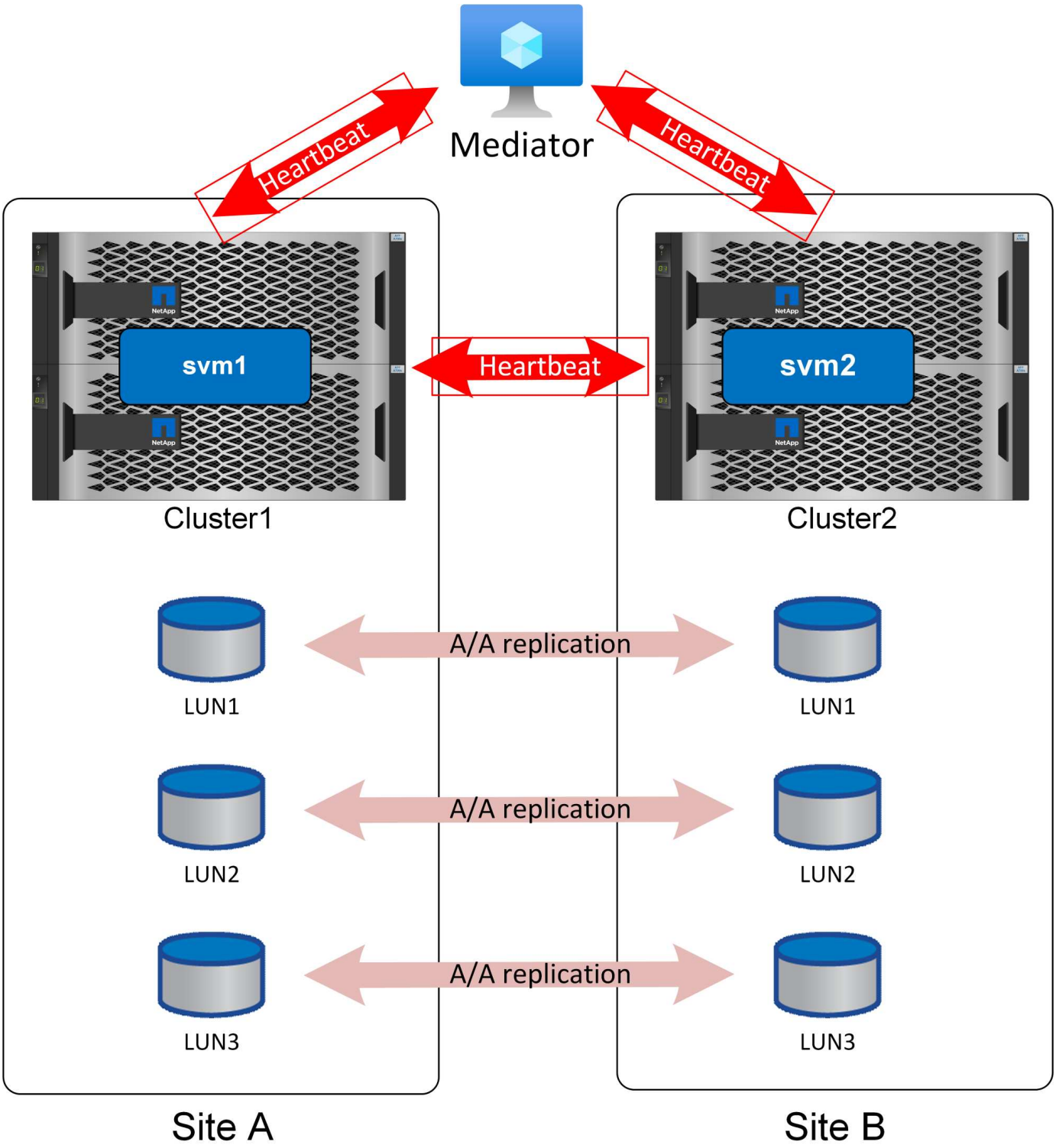
ONTAP Mediator

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.

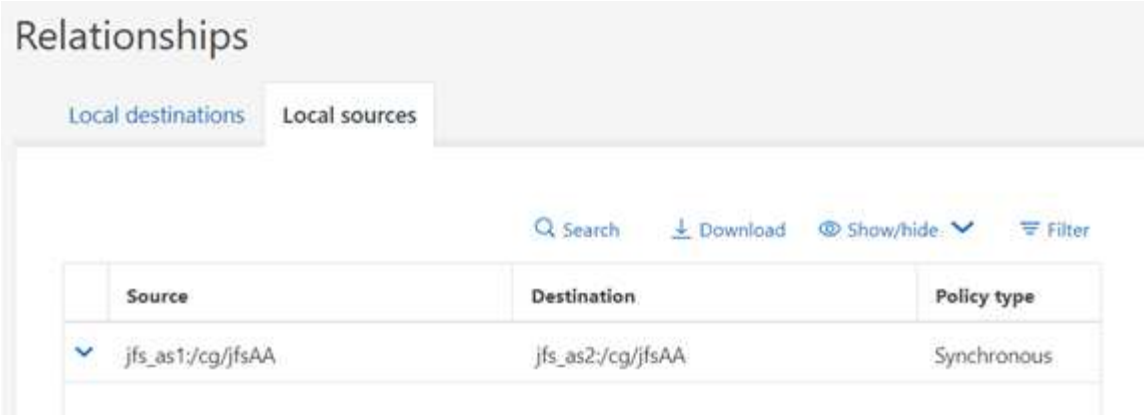
Bevorzugter Standort für SnapMirror Active Sync

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.

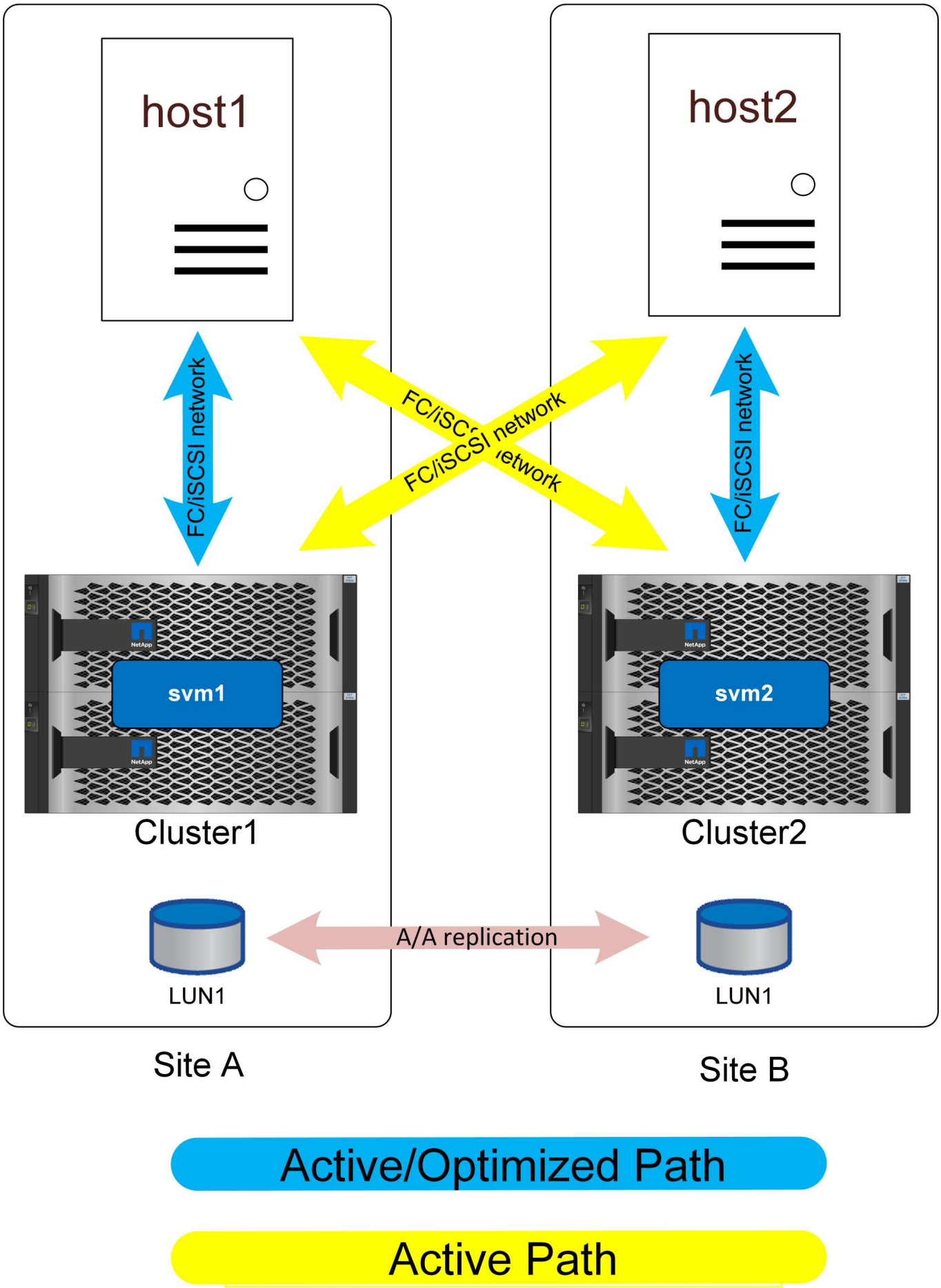
Der zweite, optionale Schritt ist die Konfiguration der Host-Nähe. Dies kontrolliert nicht den Zugriff, es steuert *Priority*.

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

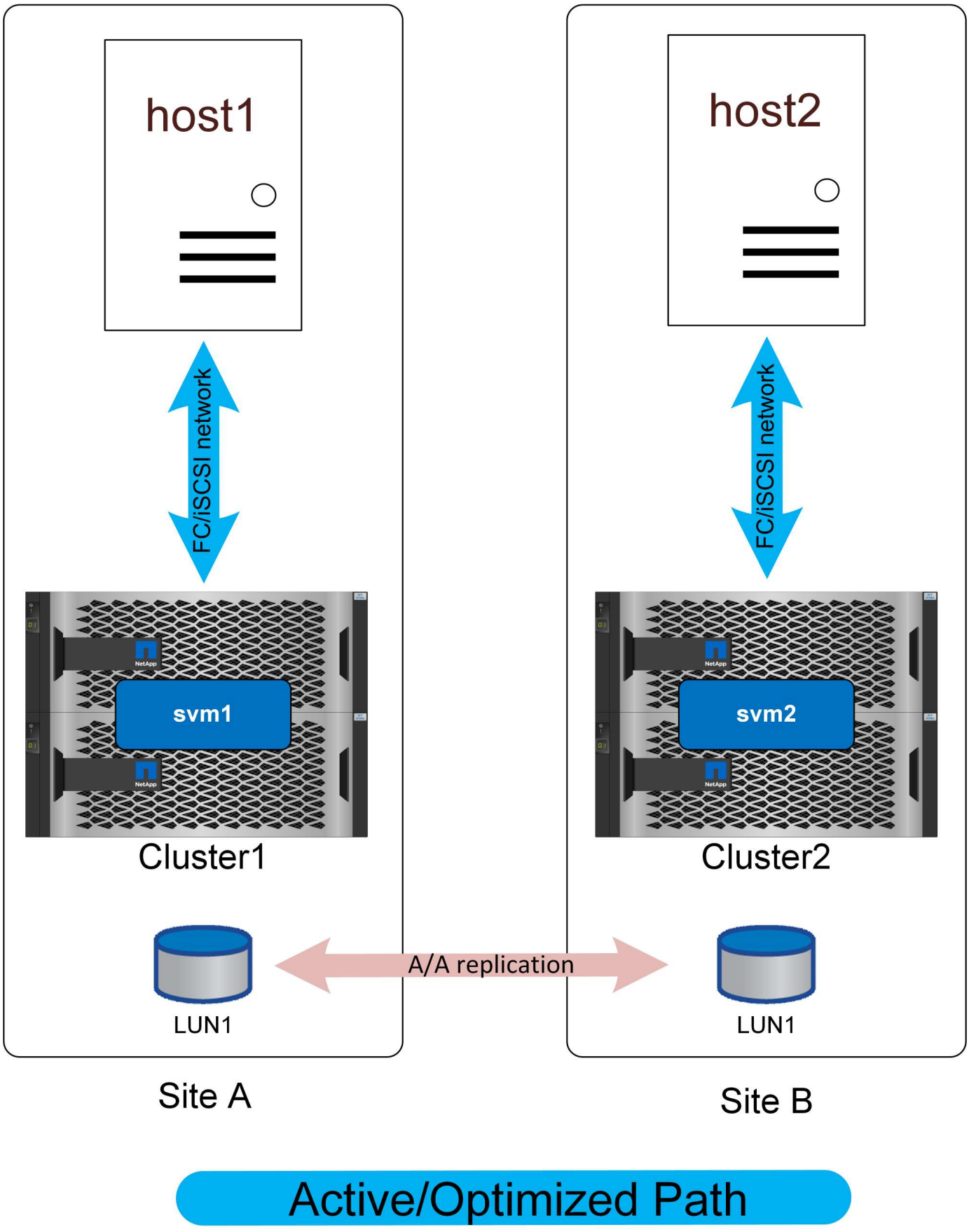
Im normalen Betrieb sind alle E/A-Vorgänge lokal. Lese- und Schreibvorgänge werden vom lokalen Speicher-Array gewartet. Schreib-I/O muss natürlich auch vom lokalen Controller auf das Remote-System repliziert werden, bevor sie bestätigt wird. Alle Lese-I/O-Vorgänge werden jedoch lokal gewartet und es 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

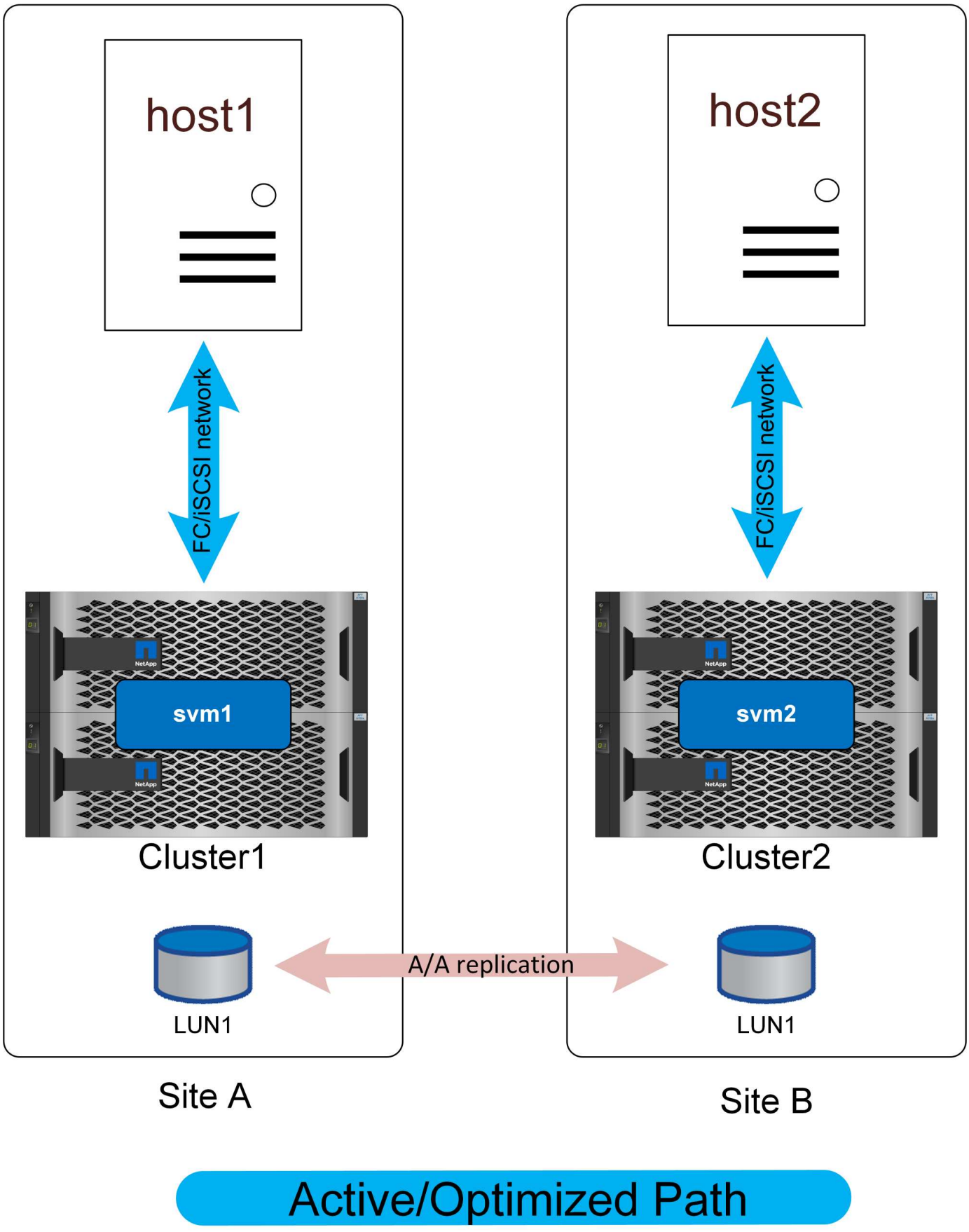
sein oder auch nicht.

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.

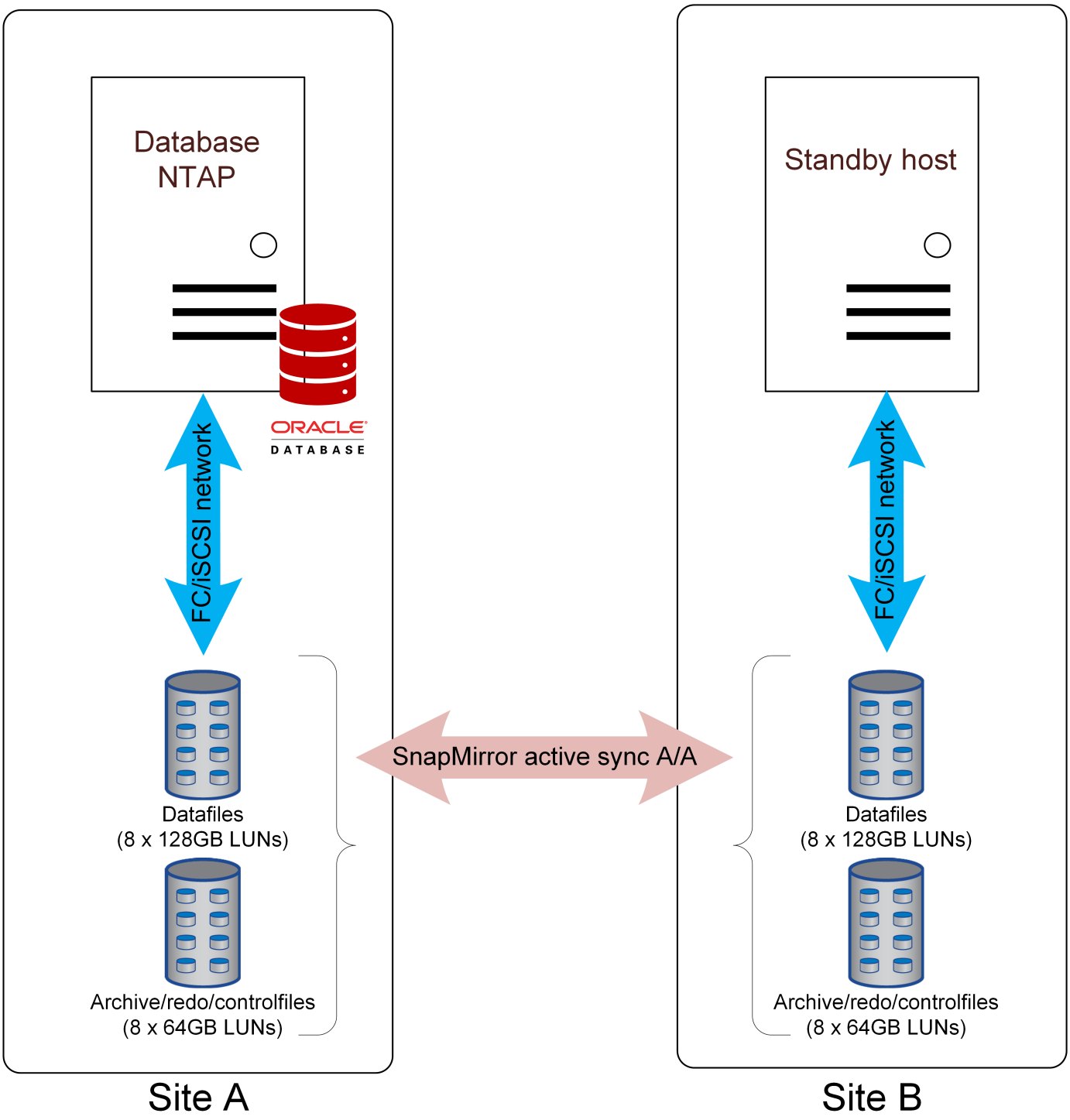
Bei einer strikeren RTO würde die grundlegende aktiv/Passiv-Automatisierung über Skripte oder Clusterware wie Pacemaker oder Ansible die Failover-Zeit verbessern. Beispielsweise könnte VMware HA konfiguriert werden, um den VM-Ausfall am primären Standort zu erkennen und die VM am Remote-Standort zu aktivieren.

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.

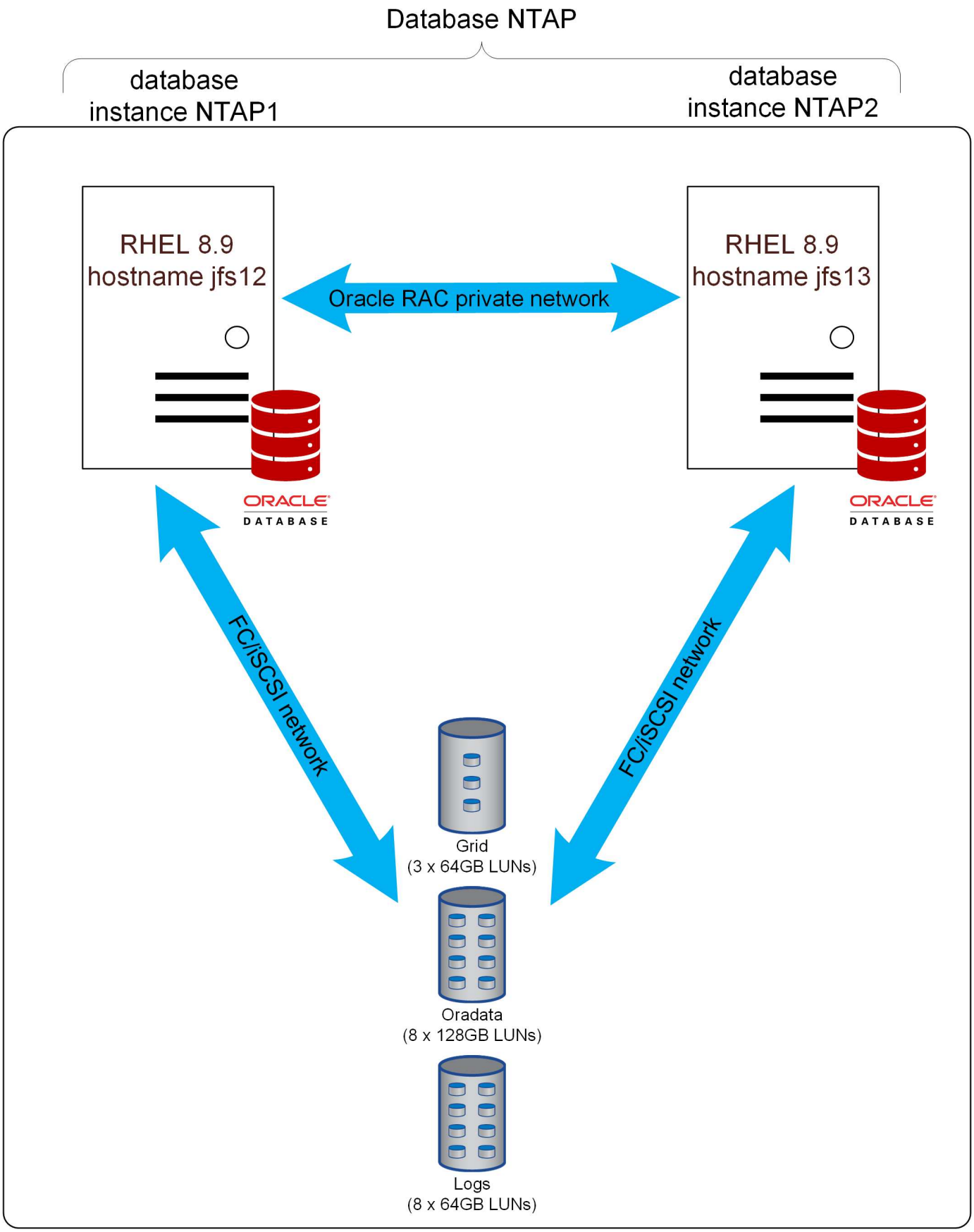
Oracle Extended RAC

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

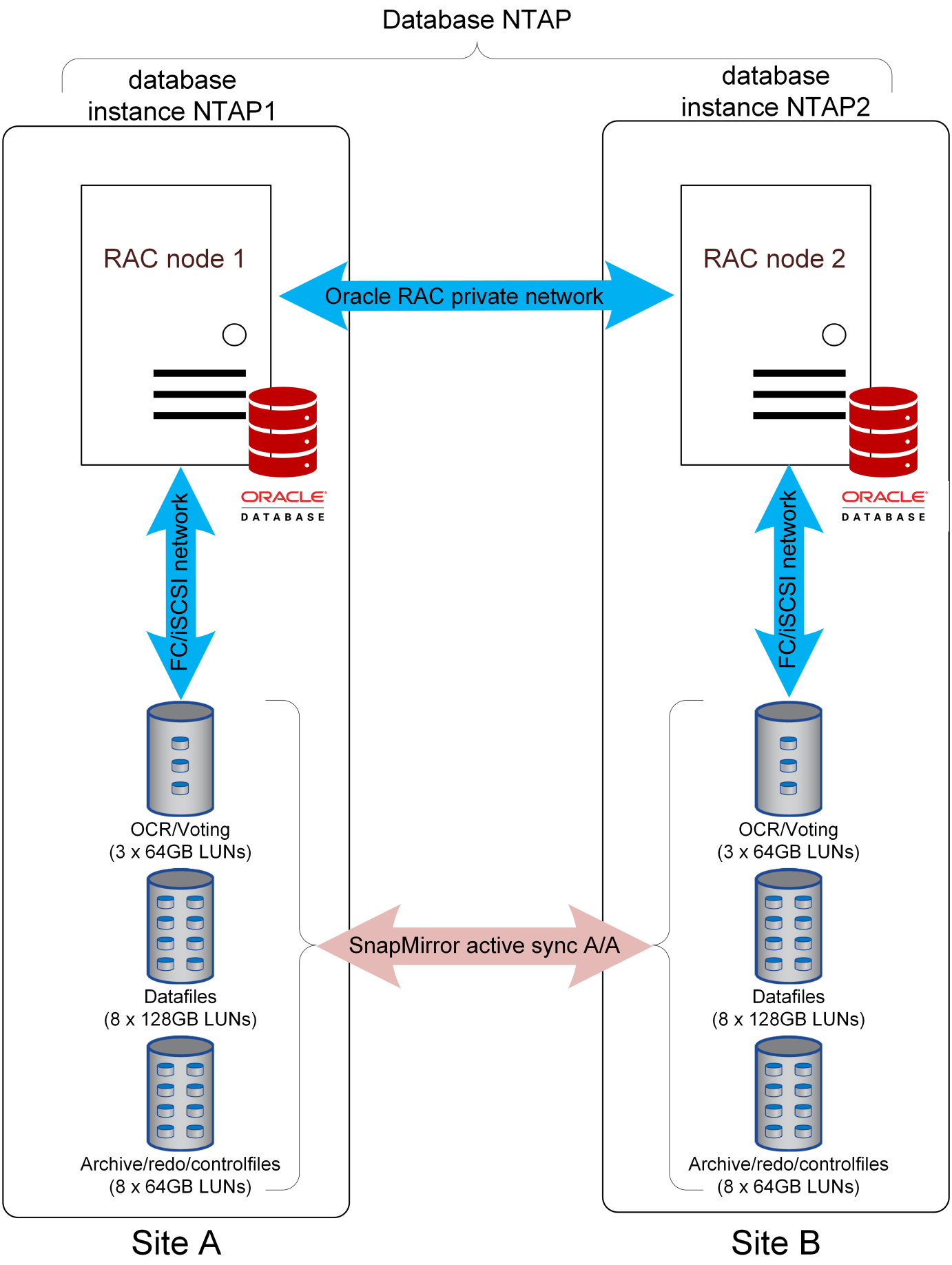
Die wichtigste Überlegung bei Oracle RAC on SnapMirror Active Sync ist, ob ein einheitlicher oder nicht einheitlicher Zugriff verwendet werden soll.

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
ORACLE DATABASE
ORACLE DATABASE
FC/iSCSI network
FC/iSCSI network
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

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

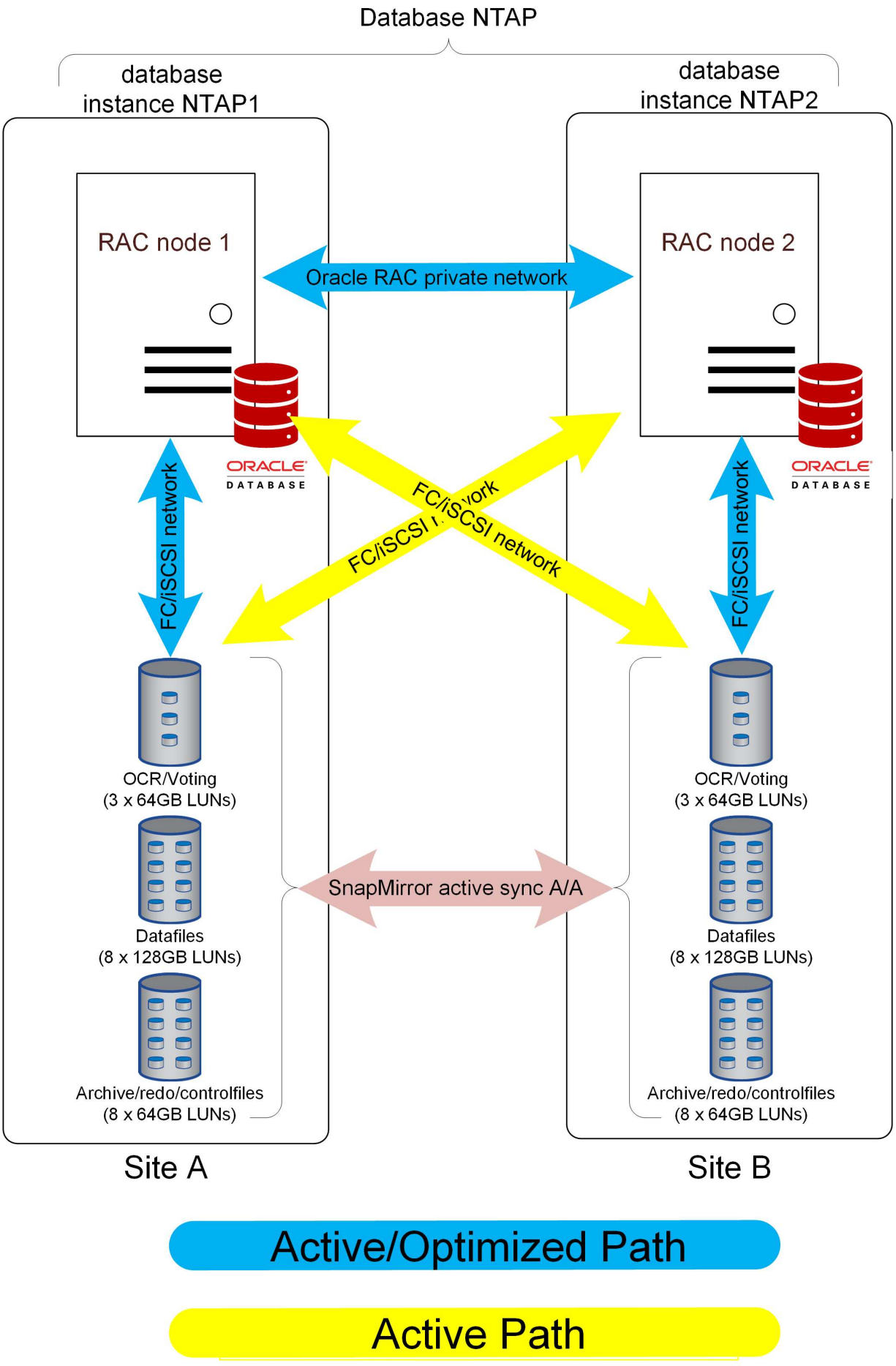
Für einen einheitlichen Zugriff muss das SAN standortübergreifend erweitert werden. Der Hauptvorteil besteht darin, dass der Verlust eines Storage-Systems nicht zum Verlust einer Datenbankinstanz führt. Stattdessen würde dies zu einer Änderung des Multipathing führen, in der Pfade derzeit verwendet werden.

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

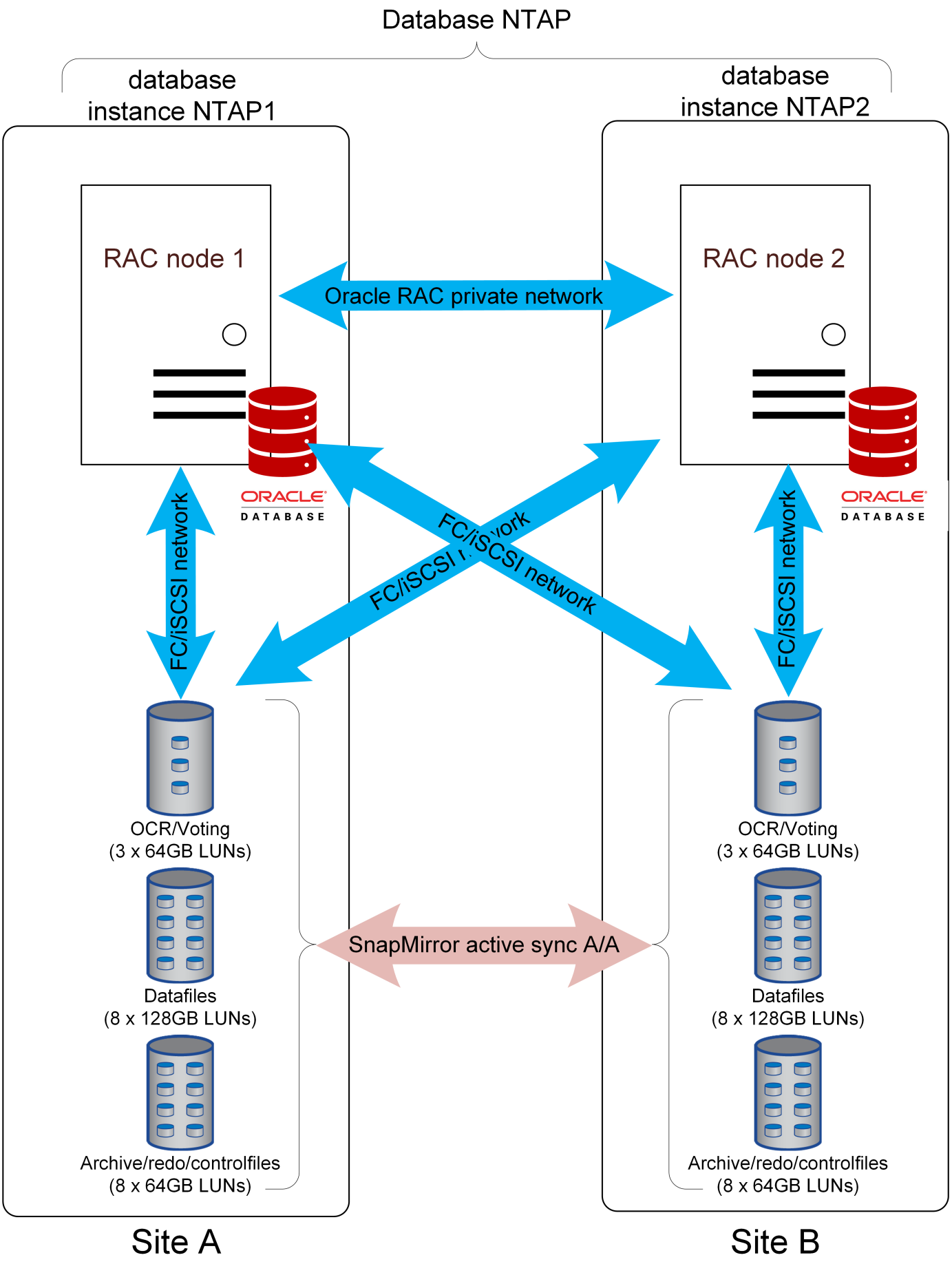
Bei erheblichen Latenzzeiten zwischen den Standorten können AFF Systeme mit Host-Näherungseinstellungen konfiguriert werden. So kann jedes Speichersystem erkennen, welche Hosts lokal und welche Remote sind, und Pfadprioritäten entsprechend zuweisen.


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

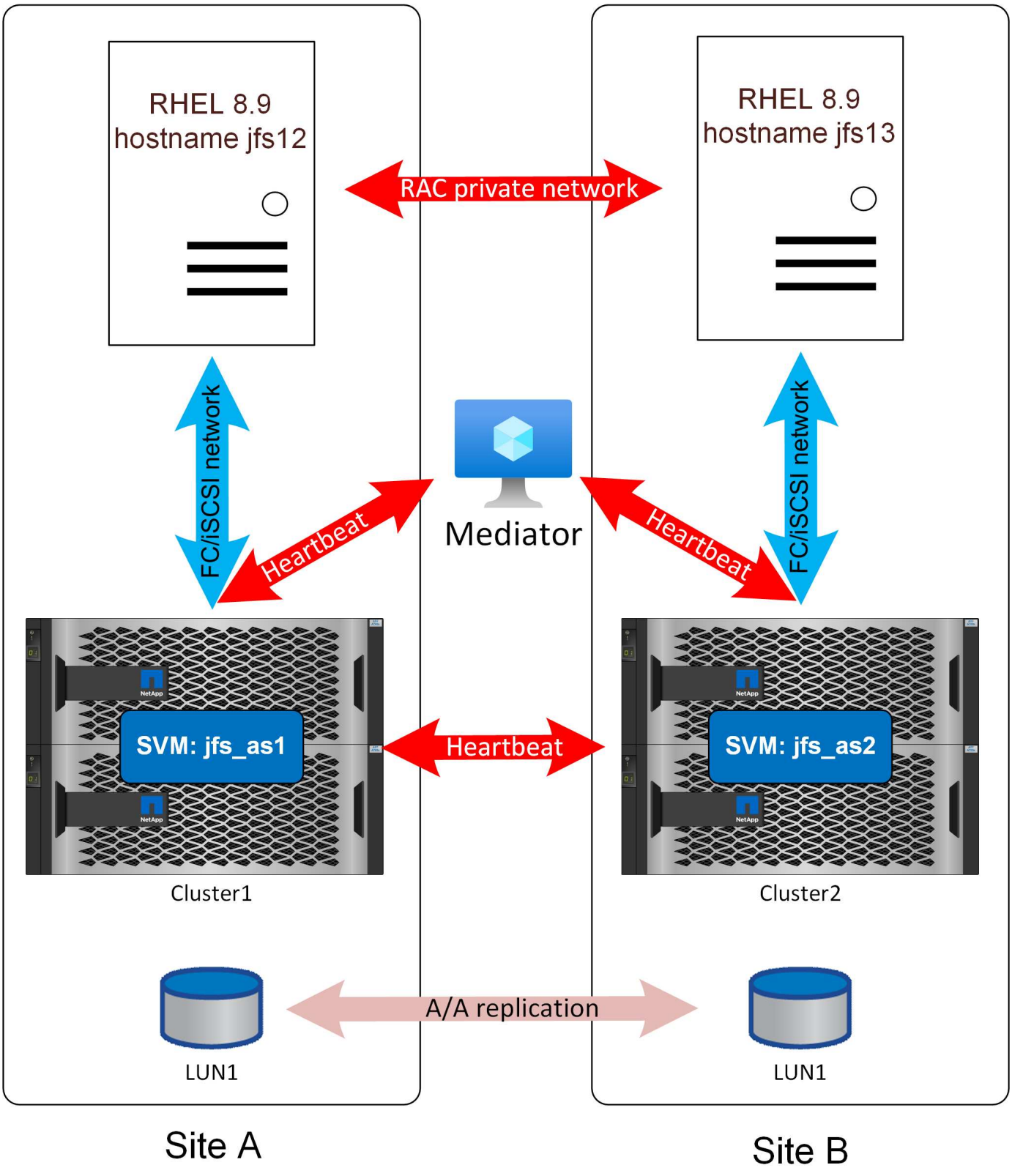
Das Ändern des bevorzugten Standorts erfordert eine einfache Bedienung. I/O-Vorgänge werden für eine oder zwei Sekunden angehalten, da zwischen den Clustern die Berechtigung für das Replikationsverhalten wechselt, die E/A-Vorgänge sind jedoch ansonsten nicht betroffen.

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.

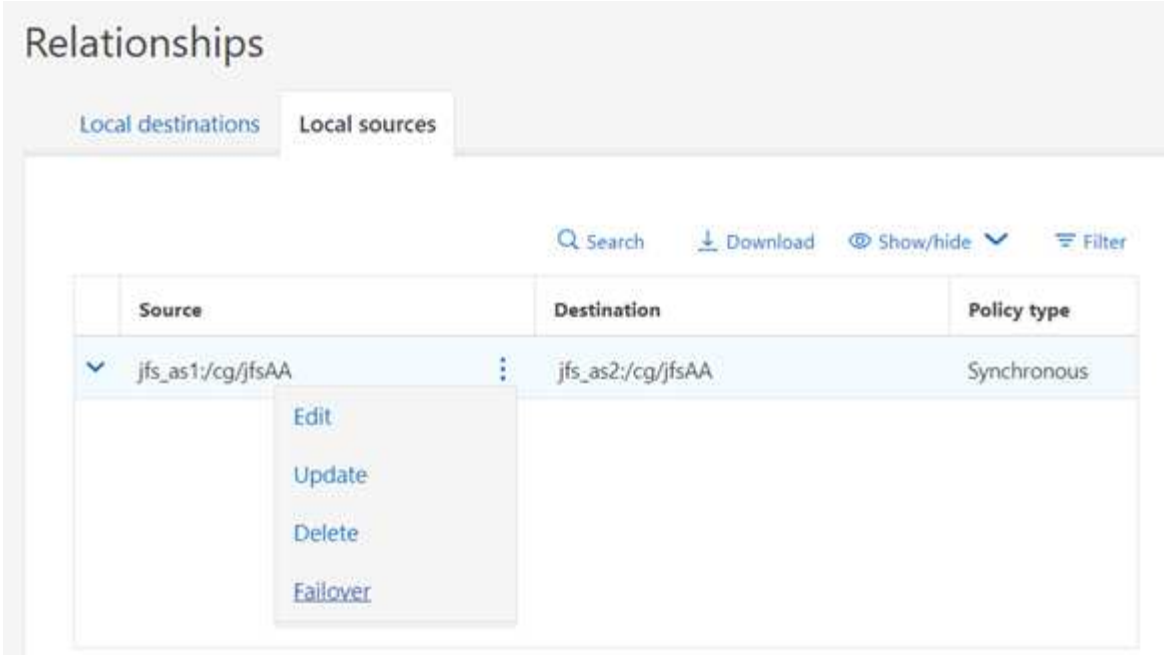
Manueller Failover

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
```

Migration der Oracle Datenbank

Überblick

Die Nutzung der Leistungsfähigkeit einer neuen Storage-Plattform erfordert zwingend die Speicherung der Daten in dem neuen Storage-System. Mit ONTAP ist der Migrationsprozess denkbar einfach: Migrationen und Upgrades von ONTAP zu ONTAP, Import fremder LUNs und Verfahren für die direkte Nutzung des Host-Betriebssystems oder der Oracle Datenbanksoftware.

Diese Dokumentation ersetzt den bereits veröffentlichten technischen Bericht *TR-4534: Migration von Oracle-Datenbanken zu NetApp-Speichersystemen*

Im Falle eines neuen Datenbankprojekts ist dies kein Problem, da die Datenbank- und Anwendungsumgebungen eingerichtet sind. Die Migration stellt Unternehmen jedoch vor besondere Herausforderungen, was die Unterbrechung des Geschäftsbetriebs, den für den Abschluss der Migration

erforderlichen Zeitaufwand, die erforderlichen Fachkompetenzen und die Risikominimierung angeht.

Skripte

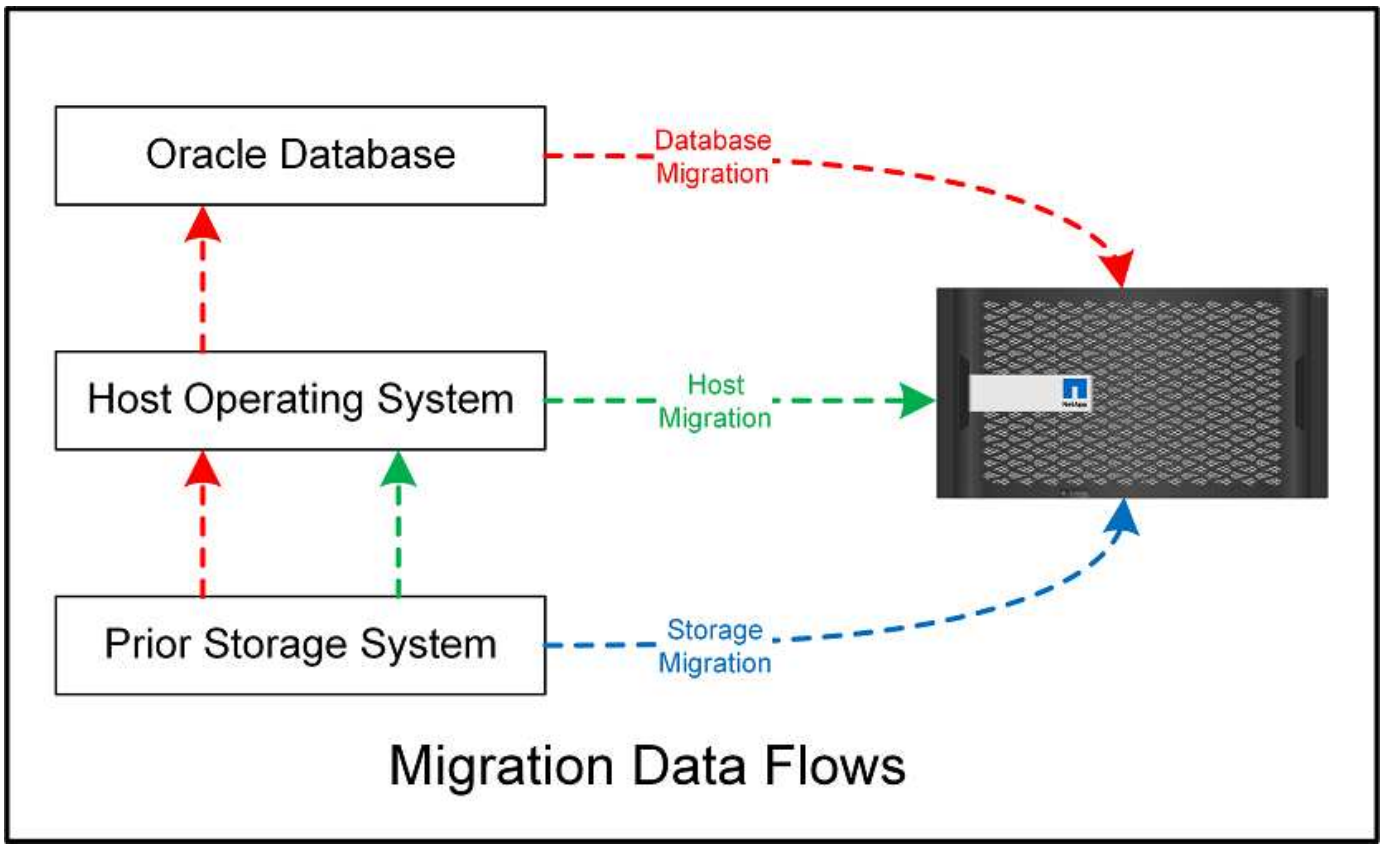
In dieser Dokumentation sind Beispielskripte enthalten. Diese Skripte bieten Beispielmethoden zur Automatisierung verschiedener Aspekte der Migration, um das Risiko von Benutzerfehlern zu verringern. Die Skripte können die Gesamtanforderungen an die für eine Migration verantwortlichen IT-Mitarbeiter verringern und den Gesamtprozess beschleunigen. Diese Skripte stammen alle aus Migrationsprojekten, die von NetApp Professional Services und NetApp-Partnern durchgeführt werden. Beispiele für deren Verwendung sind in dieser Dokumentation aufgeführt.

Migrationsplanung

Die Oracle-Datenmigration kann auf einer der drei Ebenen erfolgen: Der Datenbank, dem Host oder dem Storage Array.

Die Unterschiede liegen darin, welche Komponente der Gesamtlösung für das Verschieben von Daten verantwortlich ist: Die Datenbank, das Host-Betriebssystem oder das Speichersystem.

Die Abbildung unten zeigt ein Beispiel für die Migrationsebenen und den Datenfluss. Bei einer Migration auf Datenbankebene werden die Daten vom ursprünglichen Storage-System über die Host- und Datenbankschichten in die neue Umgebung verschoben. Die Migration auf Host-Ebene ist ähnlich, aber die Daten durchlaufen nicht die Applikationsebene und werden stattdessen mithilfe von Host-Prozessen an den neuen Speicherort geschrieben. Und schließlich ist bei der Migration auf Storage-Ebene ein Array wie ein NetApp FAS System für die Datenverschiebung verantwortlich.



Eine Migration auf Datenbankebene bezieht sich im Allgemeinen auf die Verwendung von Oracle-Protokollversand über eine Standby-Datenbank, um eine Migration auf der Oracle-Ebene durchzuführen. Migrationen auf Host-Ebene werden mithilfe der nativen Funktionen der Konfiguration des Host-

Betriebssystems durchgeführt. Diese Konfiguration umfasst Dateikopiervorgänge mit Befehlen wie CP, tar und Oracle Recovery Manager (RMAN) oder mit einem Logical Volume Manager (LVM) zur Verlagerung der zugrunde liegenden Bytes eines Dateisystems. Oracle Automatic Storage Management (ASM) wird als Funktion auf Hostebene kategorisiert, da sie unter der Ebene der Datenbankanwendung ausgeführt wird. ASM tritt an die Stelle des üblichen Logical Volume Managers auf einem Host. Und schließlich können Daten auf Storage-Array-Ebene migriert werden, d. h. unter der Ebene des Betriebssystems.

Überlegungen zur Planung

Die beste Option für die Migration hängt von einer Kombination verschiedener Faktoren ab: Vom Umfang der zu migrierenden Umgebung, der Notwendigkeit, Ausfallzeiten zu vermeiden, und dem für die Migration erforderlichen Gesamtaufwand. Große Datenbanken erfordern offensichtlich mehr Zeit und Aufwand für die Migration, aber die Komplexität einer solchen Migration ist minimal. Kleine Datenbanken können schnell migriert werden. Wenn jedoch Tausende migriert werden müssen, kann das Ausmaß des Aufwands zu Komplikationen führen. Und je größer die Datenbank ist, desto wahrscheinlicher ist sie, dass sie geschäftskritisch ist. Dies führt dazu, dass Ausfallzeiten minimiert werden müssen, während gleichzeitig ein Back-Out-Pfad beibehalten wird.

Im Folgenden werden einige Überlegungen zur Planung einer Migrationsstrategie erörtert.

Datengröße

Die Größe der zu migrierenden Datenbanken wirkt sich natürlich auf die Migrationsplanung aus. Die Größe wirkt sich jedoch nicht unbedingt auf die Umstellungszeit aus. Wenn große Datenmengen migriert werden müssen, kommt es in erster Linie auf die Bandbreite an. Kopiervorgänge werden in der Regel mit effizienten sequenziellen I/O-Vorgängen ausgeführt Gehen Sie bei Kopiervorgängen von einer Auslastung der verfügbaren Netzwerkbandbreite von 50 % aus. Ein 8-GB-FC-Port kann theoretisch etwa 800 Mbit/s übertragen. Bei einer Auslastung von 50 % kann eine Datenbank mit einer Geschwindigkeit von etwa 400 Mbps kopiert werden. Somit kann eine 10-TB-Datenbank innerhalb von etwa sieben Stunden kopiert werden.

Eine Migration über größere Entfernungen erfordert in der Regel einen kreativen Ansatz, wie der Protokollversand-Prozess in erläutert "Online-Datendatei verschieben". IP-Netzwerke über große Entfernungen verfügen selten überall in der Nähe von LAN- oder SAN-Geschwindigkeiten über eine entsprechende Bandbreite. In einem Fall unterstützte NetApp die Fernmigration einer 220 TB großen Datenbank mit sehr hohen Archiv- Log-Generierungsraten. Der gewählte Ansatz für die Datenübertragung war der tägliche Versand von Bändern, da diese Methode die maximal mögliche Bandbreite bot.

Anzahl der Datenbanken

In vielen Fällen liegt das Problem beim Verschieben großer Datenmengen nicht in der Datengröße, sondern in der Komplexität der Konfiguration, die die Datenbank unterstützt. Wenn man einfach nur weiß, dass 50 TB Datenbanken migriert werden müssen, reicht das nicht aus. Dabei kann es sich um eine einzelne 50 TB große geschäftskritische Datenbank, eine Sammlung von 4, 000 älteren Datenbanken oder eine Kombination aus Produktions- und nicht produktiven Daten handeln. In manchen Fällen besteht ein Großteil der Daten aus Klonen einer Quelldatenbank. Diese Klone müssen überhaupt nicht migriert werden, da sie einfach wiederhergestellt werden können. Dies gilt insbesondere dann, wenn die neue Architektur die Nutzung von NetApp FlexClone Volumes ermöglicht.

Für die Migrationsplanung müssen Sie verstehen, wie viele Datenbanken im Umfang enthalten sind und wie sie priorisiert werden müssen. Mit zunehmender Anzahl an Datenbanken ist die bevorzugte Migrationsoption in der Regel im Stack immer geringer. So kann zum Beispiel das Kopieren einer einzelnen Datenbank mit RMAN problemlos und bei einem kurzen Ausfall durchgeführt werden. Dies ist die Replizierung auf Host-Ebene.

Wenn es 50 Datenbanken gibt, kann es einfacher sein, die Einrichtung einer neuen Dateisystemstruktur zu vermeiden, um eine RMAN-Kopie zu erhalten und stattdessen die Daten an die Stelle zu verschieben. Dies

kann durch hostbasierte LVM-Migration erfolgen, um Daten von alten LUNs auf neue LUNs zu verschieben. Dadurch wird die Verantwortung vom Datenbankadministratorteam (DBA) an das Betriebssystemteam übertragen und die Daten werden in Bezug auf die Datenbank transparent migriert. Die Dateisystemkonfiguration ist unverändert.

Wenn schließlich 500 Datenbanken von 200 Servern migriert werden müssen, können speicherbasierte Optionen wie die ONTAP Funktion zum Importieren fremder LUNs (Foreign LUN Import, FLI) verwendet werden, um eine direkte Migration der LUNs durchzuführen.

Anforderungen neu architekturgerecht

In der Regel muss das Layout einer Datenbankdatei verändert werden, um die Funktionen des neuen Storage Array nutzen zu können. Dies ist jedoch nicht immer der Fall. Die Funktionen der EF-Series All-Flash-Arrays richten sich beispielsweise primär an SAN-Performance und SAN-Zuverlässigkeit. In den meisten Fällen können Datenbanken auf ein EF-Series Array migriert werden, ohne dass dabei besondere Überlegungen zum Datenlayout angestellt werden müssen. Die einzigen Anforderungen sind hohe IOPS, niedrige Latenz und robuste Zuverlässigkeit. Wenngleich es Best Practices für Faktoren wie RAID-Konfiguration oder Dynamic Disk Pools gibt, sind für EF-Series Projekte kaum nennenswerte Änderungen an der gesamten Storage-Architektur erforderlich, um diese Funktionen nutzen zu können.

Im Gegensatz dazu erfordert die Migration zu ONTAP in der Regel eine stärkere Berücksichtigung des Datenbanklayouts, um sicherzustellen, dass die endgültige Konfiguration den größtmöglichen Nutzen erzielt. ONTAP bietet für eine Datenbankumgebung bereits viele Funktionen ohne besondere Architekturanstrengungen. Am wichtigsten ist jedoch die Möglichkeit einer unterbrechungsfreien Migration auf neue Hardware, wenn die aktuelle Hardware das Ende ihres Lebenszyklus erreicht. Generell gilt: Eine Migration zu ONTAP ist die letzte Migration, die Sie durchführen müssen. Nachfolgende Hardware Upgrades werden durchgeführt und die Daten werden unterbrechungsfrei auf neue Medien migriert.

Bei einigen Planungen sind noch mehr Vorteile verfügbar. Die wichtigsten Überlegungen beziehen sich auf die Verwendung von Snapshots. Snapshots bilden die Grundlage für nahezu sofortige Backups, Restores und Klonvorgänge. Ein Beispiel für die Leistung von Snapshots ist der größte bekannte Einsatz bei einer einzigen Datenbank mit 996 TB, die auf ca. 250 LUNs auf 6 Controllern ausgeführt wird. Diese Datenbank kann innerhalb von 2 Minuten gesichert, innerhalb von 2 Minuten wiederhergestellt und innerhalb von 15 Minuten geklont werden. Zu den weiteren Vorteilen zählen die Möglichkeit, Daten im Cluster als Reaktion auf Workload-Änderungen zu verschieben, sowie die Anwendung von Quality-of-Service-Kontrollen (QoS), um in einer Umgebung mit mehreren Datenbanken eine gute und konsistente Performance zu erreichen.

Technologien wie QoS-Steuerung, Datenverlagerung, Snapshots und Klonen arbeiten in nahezu jeder Konfiguration. Allerdings ist einige Überlegungen im Allgemeinen erforderlich, um die Vorteile zu maximieren. In einigen Fällen können Änderungen am Design von Datenbank-Storage-Layouts erforderlich sein, um die Investitionen in das neue Speicher-Array zu maximieren. Solche Designänderungen können sich auf die Migrationsstrategie auswirken, da Host- oder Storage-basierte Migrationen das ursprüngliche Datenlayout replizieren. Weitere Schritte sind möglicherweise erforderlich, um die Migration abzuschließen und ein für ONTAP optimiertes Daten-Layout zu liefern. Die in dargestellten Verfahren "Oracle-Migrationsverfahren – Überblick" Und später zeigen einige der Methoden, um nicht nur eine Datenbank zu migrieren, sondern sie mit minimalem Aufwand in das optimale finale Layout zu migrieren.

Umstellungszeit

Der maximal zulässige Service-Ausfall während der Umstellung sollte ermittelt werden. Es ist ein häufiger Fehler anzunehmen, dass der gesamte Migrationsprozess zu Störungen führt. Viele Aufgaben können vor Beginn von Serviceunterbrechungen durchgeführt werden. Viele Optionen ermöglichen den Abschluss der Migration ohne Unterbrechungen oder Ausfälle. Auch wenn Unterbrechungen unvermeidbar sind, müssen Sie dennoch den maximal zulässigen Serviceausfall definieren, da die Dauer der Umstellungszeit von Prozedur zu Prozedur variiert.

Das Kopieren einer 10-TB-Datenbank dauert beispielsweise in der Regel ungefähr sieben Stunden. Wenn das Unternehmen einen Ausfall von sieben Stunden zulassen muss, ist Dateikopien eine einfache und sichere Möglichkeit für die Migration. Wenn fünf Stunden nicht akzeptabel sind, lässt sich ein einfacher Protokollversand-Prozess wie (siehe) "Oracle-Protokollversand") Kann mit minimalem Aufwand eingerichtet werden, um die Umstellungszeit auf etwa 15 Minuten zu reduzieren. Während dieser Zeit kann ein Datenbankadministrator den Prozess abschließen. Wenn 15 Minuten nicht akzeptabel sind, kann der endgültige Umstellungsprozess durch Skripting automatisiert werden, um die Umstellungszeit auf wenige Minuten zu verkürzen. Sie können eine Migration jederzeit beschleunigen, doch dies kostet Zeit und Aufwand. Die Umstellungszeitziele sollten darauf basieren, was für das Unternehmen akzeptabel ist.

Rückweg

Keine Migration ist völlig risikolos. Auch wenn die Technik einwandfrei funktioniert, besteht immer die Möglichkeit eines Anwenderfehlers. Das mit einem ausgewählten Migrationspfad verbundene Risiko muss neben den Folgen einer fehlgeschlagenen Migration berücksichtigt werden. Die transparente Online-Storage-Migrationsfunktion von Oracle ASM ist beispielsweise eines der wichtigsten Merkmale, und diese Methode ist eine der zuverlässigsten. Mit dieser Methode werden die Daten jedoch irreversibel kopiert. In dem sehr unwahrscheinlichen Fall, dass ein Problem mit ASM auftritt, gibt es keinen einfachen Rückweg. Die einzige Option besteht darin, entweder die ursprüngliche Umgebung wiederherzustellen oder die Migration mit ASM zurück zu den ursprünglichen LUNs rückgängig zu machen. Das Risiko kann durch ein Backup vom Typ Snapshot auf dem ursprünglichen Storage-System minimiert, aber nicht sogar ganz beseitigt werden, vorausgesetzt, das System ist in der Lage, einen solchen Vorgang auszuführen.

Probe

Einige Migrationsverfahren müssen vor der Ausführung vollständig überprüft werden. Eine Migration und eine Generalprobe des Umstellungsprozesses ist eine häufige Anfrage bei geschäftskritischen Datenbanken, bei denen die Migration erfolgreich sein und die Downtime minimiert werden muss. Zudem gehören auch die Anwenderakzeptanztests häufig zu den Aufgaben nach der Migration, und das gesamte System kann erst nach Abschluss der Tests in die Produktionsumgebung zurückgeführt werden.

Wenn es Bedarf an Proben gibt, können verschiedene ONTAP Funktionen den Prozess wesentlich vereinfachen. Snapshots können insbesondere eine Testumgebung zurücksetzen und schnell mehrere platzsparende Kopien einer Datenbankumgebung erstellen.

Verfahren

Überblick

Für die Oracle-Migrationsdatenbank sind zahlreiche Verfahren verfügbar. Das richtige hängt von Ihren geschäftlichen Anforderungen ab.

In vielen Fällen haben Systemadministratoren und DBAs ihre eigenen bevorzugten Methoden, um physische Volume-Daten zu verschieben, zu spiegeln und zu demirrieren oder Oracle RMAN zum Kopieren von Daten zu nutzen.

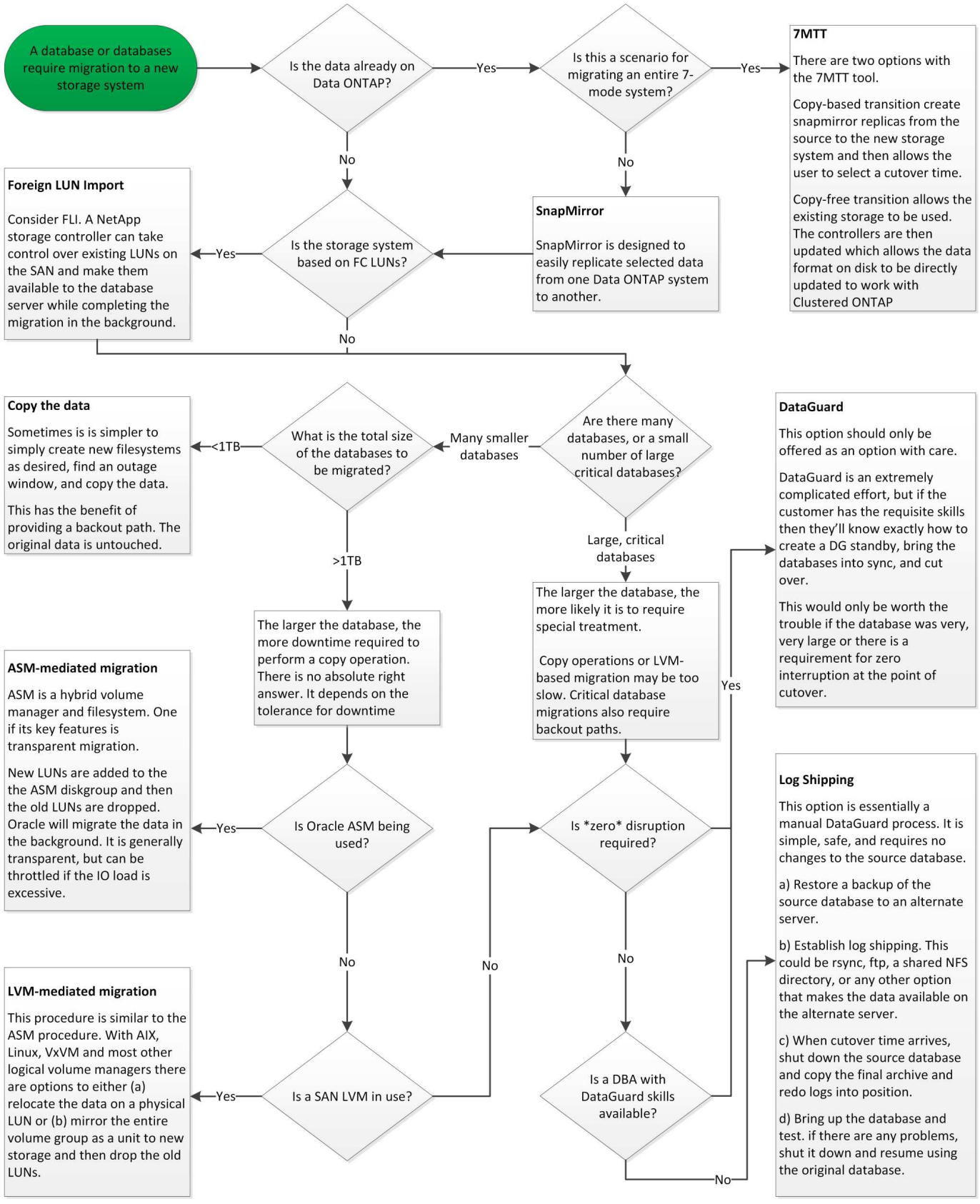
Diese Verfahren dienen in erster Linie als Orientierungshilfe für IT-Mitarbeiter, die mit einigen der verfügbaren Optionen nicht vertraut sind. Des Weiteren werden die Aufgaben, der zeitliche Bedarf und der Qualifikationsbedarf für jeden Migrationsansatz dargestellt. Dadurch können auch andere Parteien wie NetApp und Partner Professional Services oder IT-Management die Anforderungen an die einzelnen Verfahren voll einschätzen.

Es gibt keine einzigen Best Practices für die Erstellung einer Migrationsstrategie. Um einen Plan zu erstellen, müssen zunächst die Verfügbarkeitsoptionen verstanden und anschließend die Methode ausgewählt werden,

die den Anforderungen des Unternehmens am besten entspricht. Die folgende Abbildung zeigt die grundlegenden Überlegungen und typischen Schlussfolgerungen von Kunden, ist aber nicht universell auf alle Situationen anwendbar.

Ein Schritt wirft beispielsweise das Problem der Gesamtgröße der Datenbank auf. Der nächste Schritt hängt davon ab, ob die Datenbank mehr oder weniger als 1 TB umfasst. Die empfohlenen Schritte sind genau das – Empfehlungen auf der Basis typischer Kundenpraktiken. Die meisten Kunden würden nicht mit DataGuard eine kleine Datenbank kopieren, aber einige könnten. Die meisten Kunden würden aufgrund der erforderlichen Zeit nicht versuchen, eine 50 TB große Datenbank zu kopieren, aber einige haben möglicherweise ein ausreichend großes Wartungsfenster, um einen solchen Vorgang zu ermöglichen.

Das folgende Flussdiagramm zeigt die Arten von Überlegungen, welche Migrationspfade am besten geeignet sind. Sie können mit der rechten Maustaste auf das Bild klicken und es in einer neuen Registerkarte öffnen, um die Lesbarkeit zu verbessern.



.

Online-Datendatei verschieben

Bei Oracle 12cR1 und höher kann eine Datendatei verschoben werden, während die Datenbank online bleibt. Es funktioniert außerdem zwischen verschiedenen Dateisystemtypen. Eine Datendatei kann beispielsweise

von einem xfs-Dateisystem in ASM verschoben werden. Diese Methode wird im Allgemeinen nicht in der Größenordnung verwendet, da die Anzahl der erforderlichen individuellen Datendateiverschiebungsvorgänge erforderlich wäre. Es ist jedoch eine Option, die es sich lohnt, bei kleineren Datenbanken mit weniger Datendateien in Betracht zu ziehen.

Darüber hinaus ist das einfache Verschieben einer Datendatei eine gute Option für die Migration von Teilen vorhandener Datenbanken. Beispielsweise können weniger aktive Datendateien auf kostengünstigeren Storage verschoben werden, beispielsweise auf ein FabricPool Volume, mit dem ungenutzte Blöcke im Objektspeicher gespeichert werden können.

Migration auf Datenbankebene

Die Migration auf Datenbankebene bedeutet, dass die Datenbank Daten verschieben kann. Konkret bedeutet dies Protokollversand. Technologien wie RMAN und ASM sind Oracle Produkte. Im Rahmen der Migration arbeiten sie jedoch auf Hostebene, wo sie Dateien kopieren und Volumes managen.

Protokollversand

Die Grundlage für die Migration auf Datenbankebene ist das Oracle Archivprotokoll, das ein Protokoll der Änderungen an der Datenbank enthält. Meistens ist ein Archivprotokoll Bestandteil einer Backup- und Recovery-Strategie. Der Recovery-Prozess beginnt mit der Wiederherstellung einer Datenbank und dann mit der Wiedergabe eines oder mehrerer Archivprotokolle, um die Datenbank in den gewünschten Zustand zu bringen. Mit derselben Basistechnologie kann eine Migration mit nur minimaler bis keiner Unterbrechung des Betriebs durchgeführt werden. Noch wichtiger ist, dass diese Technologie die Migration ermöglicht und gleichzeitig die ursprüngliche Datenbank unberührt lässt. Dabei wird ein Back-Out-Pfad beibehalten.

Der Migrationsprozess beginnt mit der Wiederherstellung eines Datenbank-Backups auf einem sekundären Server. Dies kann auf unterschiedliche Weise erfolgen, doch die meisten Kunden verwenden ihre normale Backup-Applikation, um die Datendateien wiederherzustellen. Nachdem die Datendateien wiederhergestellt sind, legen Benutzer eine Methode für den Protokollversand fest. Das Ziel besteht darin, einen konstanten Feed von Archivprotokollen zu erstellen, die von der primären Datenbank generiert werden, und diese in der wiederhergestellten Datenbank wiederzugeben, um sie nahe am selben Status zu halten. Wenn die Umstellung ankommt, wird die Quelldatenbank vollständig heruntergefahren und die letzten Archivprotokolle sowie in einigen Fällen die Wiederherstellungsprotokolle kopiert und wiedergegeben. Es ist wichtig, dass die Wiederherstellungsprotokolle auch berücksichtigt werden, da sie einige der letzten abgeschlossenen Transaktionen enthalten können.

Nachdem diese Protokolle übertragen und wiedergegeben wurden, sind beide Datenbanken konsistent. Jetzt führen die meisten Kunden einige grundlegende Tests durch. Wenn während des Migrationsprozesses Fehler auftreten, sollte die Protokollwiedergabe Fehler melden und fehlschlagen. Es ist weiterhin ratsam, einige schnelle Tests basierend auf bekannten Abfragen oder applikationsgestützten Aktivitäten durchzuführen, um zu überprüfen, ob die Konfiguration optimal ist. Es ist auch üblich, eine abschließende Testtabelle zu erstellen, bevor die ursprüngliche Datenbank heruntergefahren wird, um zu überprüfen, ob sie in der migrierten Datenbank vorhanden ist. Dieser Schritt stellt sicher, dass während der endgültigen Protokollsynchronisierung keine Fehler gemacht wurden.

Eine einfache Log-Shipping-Migration kann Out-of-Band hinsichtlich der ursprünglichen Datenbank konfiguriert werden, was dies besonders für geschäftskritische Datenbanken nützlich macht. Für die Quelldatenbank sind keine Konfigurationsänderungen erforderlich, und die Wiederherstellung und Erstkonfiguration der Migrationsumgebung haben keine Auswirkungen auf den Produktionsbetrieb. Nachdem der Protokollversand konfiguriert wurde, werden einige I/O-Anforderungen an die Produktionsserver gestellt. Der Protokollversand besteht jedoch aus einfachen sequenziellen Lesevorgängen in den Archivprotokollen, was sich wahrscheinlich nicht auf die Performance der Produktionsdatenbank auswirken wird.

Der Protokollversand hat sich besonders für große Entfernungen bei Migrationen mit hohen Änderungsraten

bewährt. In einer Instanz wurde eine einzelne 220 TB Datenbank an einen etwa 500 Meilen entfernten neuen Standort migriert. Die Änderungsrate war extrem hoch und Sicherheitsbeschränkungen verhinderten die Nutzung einer Netzwerkverbindung. Der Protokollversand wurde durch Tape und Kurier durchgeführt. Eine Kopie der Quelldatenbank wurde zunächst mithilfe der unten beschriebenen Verfahren wiederhergestellt. Die Protokolle wurden dann wöchentlich per Kurier bis zum Zeitpunkt der Umstellung versendet, als die endgültigen Tapes zugestellt wurden und die Protokolle auf die Replikatdatenbank angewendet wurden.

Oracle DataGuard

In einigen Fällen ist eine vollständige DataGuard Umgebung gerechtfertigt. Es ist falsch, den Begriff DataGuard zu verwenden, um auf eine Protokollversendungs- oder Standby-Datenbankkonfiguration zu verweisen. Oracle DataGuard ist ein umfassendes Framework für das Management der Datenbankreplikation, es handelt sich jedoch nicht um eine Replizierungstechnologie. Der Hauptvorteil einer kompletten DataGuard-Umgebung bei einer Migration ist das transparente Umschalten von einer Datenbank zur anderen. DataGuard ermöglicht außerdem ein transparentes Switchover zurück zur Originaldatenbank, falls ein Problem erkannt wird, beispielsweise ein Problem mit der Performance oder der Netzwerkkonnektivität in der neuen Umgebung. Eine vollständig konfigurierte DataGuard-Umgebung erfordert nicht nur die Konfiguration der Datenbankschicht, sondern auch der Applikationen, damit Applikationen eine Änderung am primären Datenbankstandort erkennen können. Im Allgemeinen ist es nicht notwendig, eine Migration mit DataGuard durchzuführen, aber einige Kunden haben intern umfangreiche DataGuard-Kenntnisse und verlassen sich bei Migrationsaufgaben bereits auf diese.

Neuarchitektur

Wie bereits erläutert, erfordert die Nutzung der erweiterten Funktionen von Storage Arrays manchmal eine Änderung des Datenbank-Layouts. Darüber hinaus verändert eine Änderung des Storage-Protokolls, wie etwa das Wechsel von ASM zu einem NFS Filesystem, zwangsläufig das Filesystem-Layout.

Einer der Hauptvorteile von Protokollversandmethoden, einschließlich DataGuard, besteht darin, dass das Replizierungsziel nicht mit der Quelle übereinstimmen muss. Bei der Migration von ASM zu einem normalen Dateisystem oder umgekehrt gibt es keine Probleme mit der Verwendung eines Protokollversandansatzes. Das genaue Layout der Datendateien kann am Ziel geändert werden, um die Verwendung der Pluggable Database (PDB)-Technologie zu optimieren oder QoS-Kontrollen für bestimmte Dateien selektiv festzulegen. Mit anderen Worten: Ein Migrationsprozess auf der Basis des Protokollversand ermöglicht Ihnen eine einfache und sichere Optimierung des Datenbank-Storage-Layouts.

Server-Ressourcen

Eine Einschränkung für die Migration auf Datenbankebene besteht in der Notwendigkeit eines zweiten Servers. Dieser zweite Server kann auf zwei Arten verwendet werden:

1. Sie können den zweiten Server als permanentes neues Zuhause für die Datenbank verwenden.
2. Sie können den zweiten Server als temporären Staging-Server verwenden. Nachdem die Datenmigration zum neuen Storage-Array abgeschlossen und getestet wurde, werden die LUN- oder NFS-Dateisysteme vom Staging-Server getrennt und mit dem ursprünglichen Server verbunden.

Die erste Option ist die einfachste, aber in sehr großen Umgebungen, die sehr leistungsstarke Server erfordern, ist die Verwendung möglicherweise nicht möglich. Die zweite Option erfordert zusätzliche Arbeit, um die Dateisysteme wieder an den ursprünglichen Speicherort zu verschieben. Es kann sich um eine einfache Operation handelt, bei der NFS als Storage-Protokoll verwendet wird, da die File-Systeme vom Staging-Server abgehängt und dann wieder auf dem ursprünglichen Server gemountet werden können.

Blockbasierte Dateisysteme erfordern eine zusätzliche Arbeitsleistung für die Aktualisierung von FC-Zoning oder iSCSI-Initiatoren. Bei den meisten logischen Volume-Managern (einschließlich ASM) werden die LUNs

automatisch erkannt und online geschaltet, nachdem sie auf dem ursprünglichen Server verfügbar gemacht wurden. Einige Dateisystem- und LVM-Implementierungen erfordern jedoch möglicherweise mehr Arbeit für den Export und Import der Daten. Die genaue Vorgehensweise kann variieren, es ist jedoch im Allgemeinen einfach, ein einfaches, wiederholbares Verfahren einzurichten, um die Migration abzuschließen und die Daten auf dem ursprünglichen Server wiederherzustellen.

Es ist zwar möglich, einen Protokollversand einzurichten und eine Datenbank in einer einzigen Server-Umgebung zu replizieren, aber die neue Instanz muss eine andere Prozess-SID haben, um die Protokolle wiederzugeben. Es ist möglich, die Datenbank vorübergehend unter einem anderen Satz von Prozess-IDs mit einer anderen SID zu erstellen und später zu ändern. Dies kann jedoch zu vielen komplizierten Management-Aktivitäten und einem Risiko von Benutzerfehlern führen.

Migration auf Host-Ebene

Bei der Migration von Daten auf Hostebene müssen das Host-Betriebssystem und die zugehörigen Dienstprogramme zum Abschluss der Migration verwendet werden. Dieser Prozess umfasst alle Utilitys zum Kopieren von Daten, darunter Oracle RMAN und Oracle ASM.

Kopieren von Daten

Der Wert einer einfachen Kopieroperation sollte nicht unterschätzt werden. Moderne Netzwerkinfrastrukturen können Daten in Gigabytes pro Sekunde verschieben und Dateikopievorgänge basieren auf effizienten sequenziellen Lese- und Schreib-I/O. Im Vergleich zum Protokollversand lassen sich mehr Unterbrechungen durch Host-Kopien vermeiden, doch bei einer Migration handelt es sich nicht nur um die Datenverschiebung. Sie umfasst im Allgemeinen Änderungen am Netzwerk, den Neustartzeit der Datenbank und Tests nach der Migration.

Die tatsächlich zum Kopieren der Daten benötigte Zeit ist möglicherweise nicht signifikant. Darüber hinaus behält ein Kopiervorgang einen garantierten Back-out-Pfad bei, da die Originaldaten unverändert bleiben. Sollten während des Migrationsprozesses Probleme auftreten, können die ursprünglichen Dateisysteme mit den Originaldaten wieder aktiviert werden.

Ändern Der Plattform

Replatforming bezieht sich auf eine Änderung des CPU-Typs. Wenn eine Datenbank von einer herkömmlichen Solaris-, AIX- oder HP-UX-Plattform zu x86 Linux migriert wird, müssen die Daten aufgrund von Änderungen in der CPU-Architektur neu formatiert werden. SPARC, IA64 und POWER CPUs werden als Big-Endian-Prozessoren bezeichnet, während die x86- und x86_64-Architekturen als Little-Endian bezeichnet werden. Daher werden einige Daten in Oracle-Datendateien je nach verwendetem Prozessor unterschiedlich sortiert.

In der Vergangenheit haben Kunden Daten mithilfe von DataPump plattformübergreifend repliziert. DataPump ist ein Dienstprogramm, das einen speziellen Typ des logischen Datenexports erzeugt, der schneller in die Zieldatenbank importiert werden kann. Da es eine logische Kopie der Daten erstellt, lässt DataPump die Abhängigkeiten der Prozessorabhängigkeit hinter sich. DataPump wird von einigen Kunden weiterhin für das Replatforming verwendet, aber mit Oracle 11g ist eine schnellere Option verfügbar: Plattformübergreifende transportable Tablespaces. Mit diesem Vorschub kann ein Tablespace in ein anderes endian-Format konvertiert werden. Dies ist eine physische Transformation, die eine bessere Leistung bietet als ein DataPump-Export, der physische Bytes in logische Daten konvertieren und dann zurück in physische Bytes konvertieren muss.

Eine vollständige Diskussion über DataPump und transportable Tablespaces geht über den Umfang der NetApp-Dokumentation hinaus. NetApp hat jedoch einige Empfehlungen, die auf unseren Erfahrungen basieren, die Kunden bei der Migration zu einem neuen Storage Array-Protokoll mit einer neuen CPU-Architektur unterstützt haben:

- Wenn DataPump verwendet wird, sollte die für den Abschluss der Migration erforderliche Zeit in einer Testumgebung gemessen werden. Kunden sind manchmal überrascht, wie lange sie für die Durchführung der Migration benötigen. Diese unerwartete zusätzliche Ausfallzeit kann zu Unterbrechungen führen.
- Viele Kunden glauben irrtümlicherweise, dass plattformübergreifende transportable Tablespaces keine Datenkonvertierung erfordern. Wenn eine CPU mit einem anderen Endian verwendet wird, wird ein RMAN verwendet `convert` Der Betrieb muss zuvor an den Datendateien durchgeführt werden. Dies ist kein sofortiger Vorgang. In einigen Fällen kann der Konvertierungsprozess beschleunigt werden, indem mehrere Threads auf verschiedenen Dateien arbeiten, aber der Konvertierungsprozess kann nicht vermieden werden.

Migration über Manager eines logischen Volumes

LVMs nehmen eine Gruppe von einer oder mehreren LUNs und zerteilen sie in kleine Einheiten, die im Allgemeinen als Extents bezeichnet werden. Der Pool mit Erweiterungen wird dann als Quelle verwendet, um logische Volumes zu erstellen, die im Wesentlichen virtualisiert sind. Diese Virtualisierungsebene bietet auf verschiedene Weise einen Mehrwert:

- Logische Volumes können Extents verwenden, die von mehreren LUNs stammen. Wenn ein Filesystem auf einem logischen Volume erstellt wird, können alle Performance-Funktionen aller LUNs genutzt werden. Zudem wird die gleichmäßige Auslastung aller LUNs in der Volume-Gruppe gefördert, wodurch eine besser planbare Performance erzielt wird.
- Die Größe logischer Volumes kann durch Hinzufügen und in einigen Fällen durch Entfernen von Extents geändert werden. Die Größe eines Filesystems auf einem logischen Volume ist im Allgemeinen unterbrechungsfrei.
- Logische Volumes können unterbrechungsfrei migriert werden, indem die zugrunde liegenden Extents verschoben werden.

Migration mit einer LVM funktioniert auf zwei Arten: Ein Extent verschieben oder ein Extent spiegeln/demirrieren. Bei der LVM-Migration werden effiziente sequenzielle I/O große Blöcke eingesetzt, und es entstehen nur selten Performance-Probleme. Wenn dies zu einem Problem wird, gibt es in der Regel Optionen zur Drosselung der I/O-Rate. Dadurch erhöht sich die für den Abschluss der Migration erforderliche Zeit und gleichzeitig verringert sich die I/O-Last für Host- und Speichersysteme.

Spiegel und Demirror

Einige Volume-Manager, wie AIX LVM, erlauben dem Benutzer, die Anzahl der Kopien für jedes Extent festzulegen und zu steuern, welche Geräte die einzelnen Kopien hosten. Zur Migration wird ein vorhandenes logisches Volume erstellt, die zugrunde liegenden Extents zu den neuen Volumes gespiegelt, auf eine Synchronisierung der Kopien gewartet und anschließend die alte Kopie verworfen. Wenn ein Back- Out-Pfad gewünscht wird, kann vor dem Zeitpunkt, an dem die Spiegelungskopie abgelegt wird, ein Snapshot der Originaldaten erstellt werden. Alternativ kann der Server kurz heruntergefahren werden, um die ursprünglichen LUNs zu maskieren, bevor die enthaltenen Spiegelkopien erzwungen gelöscht werden. Dabei wird eine wiederherstellbare Kopie der Daten am ursprünglichen Speicherort aufbewahrt.

Extent-Migration

Fast alle Volume-Manager erlauben die Migration von Extents, und manchmal gibt es mehrere Optionen. Beispielsweise ermöglichen einige Volume Manager einem Administrator, die einzelnen Extents für ein bestimmtes logisches Volume von altem zu neuem Storage zu verschieben. Volume-Manager wie Linux LVM2 bieten die `pvmove` Befehl, der alle Extents auf dem angegebenen LUN-Gerät auf eine neue LUN verlagert. Nach der Evakuierung der alten LUN kann sie entfernt werden.

Das primäre Risiko für den Betrieb ist das Entfernen alter, nicht genutzter LUNs aus der Konfiguration. Beim Ändern des FC-Zoning und beim Entfernen veralteter LUN-Geräte ist besonders darauf zu achten.

Oracle Automatic Storage Management

Oracle ASM ist ein kombinierter logischer Volume-Manager und ein Dateisystem. Oracle ASM erstellt eine Sammlung von LUNs, unterteilt sie in kleine Zuweisungseinheiten und präsentiert sie als einzelnes Volume, das als ASM-Festplattengruppe bezeichnet wird. ASM bietet auch die Möglichkeit, die Laufwerksgruppe durch Festlegen des Redundanzniveaus zu spiegeln. Ein Volume kann nicht gespiegelt (externe Redundanz), gespiegelt (normale Redundanz) oder dreifach gespiegelt (hohe Redundanz) werden. Bei der Konfiguration der Redundanzstufe ist darauf zu achten, dass sie nach der Erstellung nicht mehr geändert werden kann.

ASM bietet auch Dateisystemfunktionen. Obwohl das Dateisystem nicht direkt vom Host aus sichtbar ist, kann die Oracle-Datenbank Dateien und Verzeichnisse auf einer ASM-Datenträgergruppe erstellen, verschieben und löschen. Außerdem kann die Struktur mit dem Dienstprogramm asmcmd navigiert werden.

Wie bei anderen LVM-Implementierungen optimiert Oracle ASM die I/O-Performance durch Striping und Lastausgleich der I/O-Vorgänge jeder Datei über alle verfügbaren LUNs. Zweitens können die zugrunde liegenden Extents verschoben werden, um sowohl die Größenänderung der ASM-Datenträgergruppe als auch die Migration zu ermöglichen. Oracle ASM automatisiert den Prozess durch den Rebalancing-Vorgang. Neue LUNs werden einer ASM-Festplattengruppe hinzugefügt und alte LUNs werden verworfen. Dies führt zu einer Extent-Verschiebung und einem nachfolgenden Drop der evakuierten LUN aus der Festplattengruppe. Dieser Prozess ist eine der bewährtesten Migrationsmethoden, und die Zuverlässigkeit von ASM bei der Bereitstellung einer transparenten Migration ist möglicherweise das wichtigste Merkmal.

Da die Spiegelungsebene von Oracle ASM fest festgelegt ist, kann sie nicht mit der Mirror- und Demirror-Methode der Migration verwendet werden.

Migration auf Storage-Ebene

Bei der Migration auf Storage-Ebene wird die Migration sowohl unter der Applikations- als auch unter der Betriebssystemebene durchgeführt. In der Vergangenheit bedeutete dies manchmal, spezialisierte Geräte zu verwenden, auf denen LUNs auf Netzwerkebene kopiert werden konnten. Diese Funktionen finden sich jedoch jetzt nativ in ONTAP.

SnapMirror

Mit der Datenreplizierungssoftware NetApp SnapMirror erfolgt die Migration von Datenbanken zwischen NetApp Systemen nahezu universell. Der Prozess beinhaltet die Einrichtung einer Spiegelbeziehung für die zu migrierenden Volumes, um sie zu synchronisieren und dann auf das Umstellungsfenster zu warten. Wenn sie eintrifft, wird die Quelldatenbank heruntergefahren, eine letzte Aktualisierung der Spiegelung durchgeführt und die Spiegelung wird unterbrochen. Die Replikatvolumes können dann verwendet werden, indem entweder ein enthaltenes NFS-Dateisystem-Verzeichnis gemountet oder die enthaltenen LUNs ermittelt und die Datenbank gestartet wird.

Das Verschieben von Volumes innerhalb eines einzigen ONTAP Clusters gilt nicht als Migration, sondern als Routine `volume move` Betrieb. SnapMirror wird als Datenreplizierungs-Engine im Cluster eingesetzt. Dieser Prozess ist vollständig automatisiert. Es gibt keine weiteren Migrationsschritte, die durchgeführt werden müssen, wenn Attribute des Volume, wie z. B. LUN-Zuordnung oder NFS-Exportberechtigungen, mit dem Volume selbst verschoben werden. Die Standortverlagerung hat keine Unterbrechung des Host-Betriebs. In manchen Fällen muss der Netzwerkzugriff aktualisiert werden, um sicherzustellen, dass auf die neu verlagerten Daten so effizient wie möglich zugegriffen wird. Diese Aufgaben sind aber auch

unterbrechungsfrei.

Import fremder LUNs (FLI)

FLI ist eine Funktion, mit der ein Data ONTAP-System mit 8.3 oder höher eine vorhandene LUN von einem anderen Storage-Array migrieren kann. Das Verfahren ist einfach: Das ONTAP-System ist auf das bestehende Speicher-Array abgegrenzt, als ob es sich um einen anderen SAN-Host handelt. Data ONTAP übernimmt dann die Kontrolle über die gewünschten Legacy-LUNs und migriert die zugrunde liegenden Daten. Außerdem kommen bei der Migration von Daten im Importprozess die Effizienzeinstellungen des neuen Volume zum Einsatz, sodass Daten während des Migrationsprozesses inline komprimiert und dedupliziert werden können.

Die erste Implementierung von FLI in Data ONTAP 8.3 erlaubte nur Offline-Migration. Dies war ein extrem schneller Transfer, aber trotzdem bedeuteten die LUN-Daten, dass sie erst nach Abschluss der Migration verfügbar waren. Die Online-Migration wurde mit Data ONTAP 8.3 eingeführt. Diese Migration minimiert Unterbrechungen, da ONTAP während der Übertragung LUN-Daten bereitstellen kann. Während die Host-Zone neu aufgeteilt wird, um die LUNs über ONTAP zu verwenden, kommt es zu einer kurzen Unterbrechung. Sobald diese Änderungen jedoch vorgenommen werden, sind die Daten wieder verfügbar und bleiben während des gesamten Migrationsprozesses zugänglich.

Lese-I/O wird über ONTAP als Proxy übertragen, bis der Kopiervorgang abgeschlossen ist, während Schreib-I/O synchron sowohl auf die fremde als auch auf die ONTAP-LUN geschrieben wird. Die beiden LUN-Kopien werden auf diese Weise synchron gehalten, bis der Administrator eine vollständige Umstellung ausführt, die die fremde LUN freigibt und Schreibvorgänge nicht mehr repliziert.

FLI ist für den Einsatz mit FC konzipiert. Wenn jedoch ein Wechsel zu iSCSI gewünscht wird, kann die migrierte LUN nach Abschluss der Migration problemlos als iSCSI-LUN neu zugeordnet werden.

Zu den Merkmalen von FLI gehört die automatische Ausrichtungserkennung und -Einstellung. In diesem Kontext bezieht sich der Begriff „Alignment“ auf eine Partition auf einem LUN-Gerät. Für eine optimale Performance muss der I/O mit 4-KB-Blöcken abgestimmt werden. Wenn eine Partition auf einem Offset platziert wird, der kein Vielfaches von 4K ist, leidet die Performance.

Es gibt einen zweiten Aspekt der Ausrichtung, der nicht korrigiert werden kann, indem ein Partitionsoffset angepasst wird: Die Blockgröße des Dateisystems. Ein ZFS-Dateisystem beispielsweise hat in der Regel eine interne Blockgröße von 512 Byte. Andere Kunden, die AIX verwenden, haben gelegentlich jfs2-Dateisysteme mit einer 512- oder 1, 024-Byte-Blockgröße erstellt. Auch wenn das Filesystem an eine 4-KB-Grenze ausgerichtet ist, bleiben die in diesem Filesystem erstellten Dateien jedoch nicht und die Performance leidet.

FLI sollte unter diesen Umständen nicht verwendet werden. Obwohl nach der Migration auf die Daten zugegriffen werden kann, ergeben sich daraus Filesysteme mit erheblichen Performance-Einschränkungen. Grundsätzlich sollte jedes Filesystem, das einen zufälligen Überschreibvorgang auf ONTAP unterstützt, eine 4-KB-Blockgröße verwenden. Dies gilt insbesondere für Workloads wie Datenbankdateien und VDI-Implementierungen. Die Blockgröße kann mit den entsprechenden Host-Betriebssystembefehlen identifiziert werden.

Auf AIX kann beispielsweise die Blockgröße mit angezeigt werden `lsfs -q`. Mit Linux `xfs_info` Und `tune2fs` Kann für verwendet werden `xfs` Und `ext3/ext4`. Mit `zfs`, Der Befehl lautet `zdb -C`.

Der Parameter, der die Blockgröße steuert, ist `ashift` Und im Allgemeinen ist der Standardwert 9, was 2^9 oder 512 Byte bedeutet. Für eine optimale Leistung, die `ashift` Wert muss 12 (2^12=4K) sein. Dieser Wert wird zum Zeitpunkt der Erstellung des zpool gesetzt und kann nicht geändert werden, was bedeutet, dass Data zpools mit einem `ashift` Andere als 12 sollten durch Kopieren der Daten in einen neu erstellten zpool migriert werden.

Oracle ASM hat keine grundlegende Blockgröße. Die einzige Voraussetzung ist, dass die Partition, auf der die

ASM-Festplatte erstellt wird, ordnungsgemäß ausgerichtet sein muss.

7-Mode Transition Tool

Bei dem 7-Mode Transition Tool (7MTT) handelt es sich um ein Automatisierungstool zur Migration großer 7-Mode Konfigurationen zu ONTAP. Die meisten Datenbankkunden finden andere Methoden einfacher, zum Teil, da sie in der Regel ihre Umgebungen einer Datenbank nach Datenbank migrieren, anstatt den gesamten Storage-Platzbedarf zu verschieben. Zudem sind Datenbanken häufig nur ein Teil einer größeren Storage-Umgebung. Daher werden Datenbanken oft einzeln migriert und die restliche Umgebung kann mit 7MTT verschoben werden.

Es gibt eine kleine aber beträchtliche Anzahl von Kunden, die Storage-Systeme haben, die komplizierten Datenbankumgebungen gewidmet sind. Diese Umgebungen können viele Volumes, Snapshots und zahlreiche Konfigurationsdetails wie Exportberechtigungen, LUN-Initiatorgruppen, Benutzerberechtigungen und die Konfiguration des Lightweight Directory Access Protocol enthalten. In diesen Fällen können die Automatisierungsfunktionen von 7MTT die Migration vereinfachen.

7MTT kann in einem der beiden Modi ausgeführt werden:

- **Copy- Based Transition (CBT).** 7MTT mit CBT richtet SnapMirror Volumes aus einem bestehenden 7-Mode System in der neuen Umgebung ein. Nachdem die Daten synchronisiert sind, orchestriert 7MTT den Umstellungsprozess.
- **Copy- Free Transition (CFT).** 7MTT mit CFT basiert auf der in-Place Konvertierung vorhandener 7-Mode Platten-Shelfs. Es werden keine Daten kopiert und die vorhandenen Festplatten-Shelfs können wieder verwendet werden. Die vorhandene Konfiguration für Datensicherung und Storage-Effizienz bleibt erhalten.

Der primäre Unterschied zwischen diesen beiden Optionen ist der Copy-Free Transition. Er ist ein „Big-Bang“-Ansatz, bei dem alle mit dem ursprünglichen 7-Mode HA-Paar verbundenen Platten-Shelfs in die neue Umgebung verschoben werden müssen. Eine Untergruppe von Shelfs lässt sich nicht verschieben. Durch den Copy-basierten Ansatz können ausgewählte Volumes verschoben werden. Es besteht auch die Möglichkeit, dass ein längeres Umstellungsfenster mit Copy-Free Transition möglich ist, da für die Neuerstellung von Festplatten-Shelfs und die Konvertierung von Metadaten eine Verbindung erforderlich ist. Je nach Praxiserfahrung empfiehlt NetApp, für die Verlagerung und Neuverkabelung von Festplatten-Shelfs eine Stunde und für die Metadatenkonvertierung zwischen 15 Minuten und 2 Stunden zu verwenden.

Migration von Datendateien

Einzelne Oracle Datendateien können mit einem einzigen Befehl verschoben werden.

Mit dem folgenden Befehl wird beispielsweise die Datendatei IOPST.dbf aus dem Dateisystem verschoben `/oradata2` Zu Dateisystem `/oradata3`.

```
SQL> alter database move datafile  '/oradata2/NTAP/IOPS002.dbf' to
'/oradata3/NTAP/IOPS002.dbf';
Database altered.
```

Das Verschieben einer Datendatei mit dieser Methode kann langsam sein, sollte jedoch normalerweise nicht genügend I/O produzieren, um die täglichen Datenbank-Workloads zu beeinträchtigen. Im Gegensatz dazu kann die Migration über die ASM-Ausbalancierung viel schneller ablaufen, doch dies geschieht auf Kosten der Verlangsamung der gesamten Datenbank, während die Daten verschoben werden.

Die zum Verschieben von Datendateien erforderliche Zeit kann einfach gemessen werden, indem eine Test-

Datendatei erstellt und dann verschoben wird. Die verstrichene Zeit für den Vorgang wird in den Sitzungsdaten mit den folgenden Kosten aufgezeichnet:

```
SQL> set linesize 300;
SQL> select elapsed_seconds||':'||message from v$session_longops;
ELAPSED_SECONDS||':'||MESSAGE
------------------------------------------------------------------------
---------------
351:Online data file move: data file 8: 22548578304 out of 22548578304
bytes done
SQL> select bytes / 1024 / 1024 /1024 as GB from dba_data_files where
FILE_ID = 8;
        GB
----------
        21
```

In diesem Beispiel handelte es sich bei der verschobenen Datei um die Datendatei 8, deren Größe 21 GB betrug und die Migration ca. 6 Minuten in Anspruch nahm. Der erforderliche Zeitaufwand hängt natürlich von den Funktionen des Storage-Systems, des Storage-Netzwerks und der gesamten Datenbankaktivität zum Zeitpunkt der Migration ab.

Protokollversand

Ziel bei einer Migration mit Protokollversand ist, eine Kopie der ursprünglichen Datendateien an einem neuen Standort zu erstellen und anschließend eine Methode für den Versand von Änderungen in die neue Umgebung zu definieren.

Nach der Einrichtung können Protokollversand und -Wiedergabe automatisiert werden, um die Replikatdatenbank weitgehend mit der Quelle synchron zu halten. So kann beispielsweise ein Cron-Job so geplant werden, dass (a) die letzten Protokolle an den neuen Speicherort kopiert und (b) alle 15 Minuten erneut wiedergegeben werden. Dadurch sind zum Zeitpunkt der Umstellung nur minimale Unterbrechungen möglich, da maximal 15 Minuten Archiv-Logs wieder eingespielt werden müssen.

Das unten abgebildete Verfahren ist außerdem im Wesentlichen eine Datenbankklonoperation. Die gezeigte Logik ähnelt der Engine in NetApp SnapManager für Oracle (SMO) und dem NetApp SnapCenter Oracle Plug-in. Einige Kunden haben das in Skripten oder WFA Workflows angezeigte Verfahren für individuelle Klonvorgänge verwendet. Dieses Verfahren ist zwar mehr manuell als SMO oder SnapCenter, es wird jedoch immer noch ohne Skripte erstellt und die Datenmanagement-APIs in ONTAP vereinfachen den Prozess weiter.

Protokollversand – Dateisystem an Dateisystem

Dieses Beispiel zeigt die Migration einer Datenbank namens WAFFLE von einem gewöhnlichen Dateisystem zu einem anderen gewöhnlichen Dateisystem auf einem anderen Server. Es veranschaulicht auch die Verwendung von SnapMirror zum Erstellen einer schnellen Kopie der Datendateien, aber dies ist kein integraler Bestandteil des gesamten Verfahrens.

Erstellen Sie eine Datenbanksicherung

Der erste Schritt besteht darin, ein Datenbank-Backup zu erstellen. Insbesondere erfordert dieses Verfahren eine Reihe von Datendateien, die für die Wiedergabe des Archivprotokolls verwendet werden können.

Umgebung

In diesem Beispiel befindet sich die Quelldatenbank auf einem ONTAP-System. Die einfachste Methode, um ein Backup einer Datenbank zu erstellen, ist die Verwendung eines Snapshots. Die Datenbank wird für einige Sekunden in den Hot Backup-Modus versetzt, während ein `snapshot create` Der Vorgang wird auf dem Volume ausgeführt, auf dem die Datendateien gehostet werden.

```
SQL> alter database begin backup;
Database altered.
```

```
Cluster01::*> snapshot create -vserver vserver1 -volume jfsc1_oradata
hotbackup
Cluster01::*>
```

```
SQL> alter database end backup;
Database altered.
```

Das Ergebnis ist ein Snapshot auf der Festplatte, der aufgerufen wird `hotbackup` Die ein Image der Datendateien enthält, während sie sich im Hot Backup-Modus befinden. Wenn die Daten in diesem Snapshot mit den entsprechenden Archivprotokollen konsistent sind, können sie als Grundlage für eine Wiederherstellung oder einen Klon verwendet werden. In diesem Fall wird sie auf den neuen Server repliziert.

Wiederherstellung in neuer Umgebung

Das Backup muss nun in der neuen Umgebung wiederhergestellt werden. Dies kann auf verschiedene Arten erfolgen, z. B. Oracle RMAN, Wiederherstellung über eine Backup-Applikation wie NetBackup oder ein einfacher Kopiervorgang von Datendateien, die im Hot-Backup-Modus platziert wurden.

In diesem Beispiel wird SnapMirror verwendet, um das Snapshot Hot-Backup an einen neuen Speicherort zu replizieren.

1. Erstellen Sie ein neues Volume für den Empfang der Snapshot-Daten. Initialisieren Sie die Spiegelung von `jfsc1_oradata` Bis `vol_oradata`.

   ```
   Cluster01::*> volume create -vserver vserver1 -volume vol_oradata
   -aggregate data_01 -size 20g -state online -type DP -snapshot-policy
   none -policy jfsc3
   [Job 833] Job succeeded: Successful
   ```

```
Cluster01::*> snapmirror initialize -source-path vserver1:jfsc1_oradata
-destination-path vserver1:vol_oradata
Operation is queued: snapmirror initialize of destination
"vserver1:vol_oradata".
Cluster01::*> volume mount -vserver vserver1 -volume vol_oradata
-junction-path /vol_oradata
Cluster01::*>
```

2. Nachdem der Status von SnapMirror festgelegt wurde und Sie angeben, dass die Synchronisierung abgeschlossen ist, aktualisieren Sie die Spiegelung speziell auf der Grundlage des gewünschten Snapshots.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields state
source-path             destination-path        state
----------------------- ----------------------- ------------
vserver1:jfsc1_oradata vserver1:vol_oradata SnapMirrored
```

```
Cluster01::*> snapmirror update -destination-path vserver1:vol_oradata
-source-snapshot hotbackup
Operation is queued: snapmirror update of destination
"vserver1:vol_oradata".
```

3. Die erfolgreiche Synchronisierung kann durch Anzeigen des überprüft werden `newest-snapshot` Feld auf der Spiegelungslautstärke.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields newest-snapshot
source-path             destination-path        newest-snapshot
----------------------- ----------------------- ---------------
vserver1:jfsc1_oradata vserver1:vol_oradata hotbackup
```

4. Der Spiegel kann dann gebrochen werden.

```
Cluster01::> snapmirror break -destination-path vserver1:vol_oradata
Operation succeeded: snapmirror break for destination
"vserver1:vol_oradata".
Cluster01::>
```

5. Mounten Sie das neue Dateisystem.bei blockbasierten Dateisystemen variieren die genauen Verfahren je nach der verwendeten LVM. FC-Zoning oder iSCSI-Verbindungen müssen konfiguriert werden. Nachdem die Verbindung zu den LUNs hergestellt wurde, Befehle wie Linux `pvscan` Möglicherweise muss ermittelt

werden, welche Volume-Gruppen oder LUNs ordnungsgemäß konfiguriert werden müssen, damit sie von ASM erkannt werden können.

In diesem Beispiel wird ein einfaches NFS-Dateisystem verwendet. Dieses Dateisystem kann direkt gemountet werden.

```
fas8060-nfs1:/vol_oradata        19922944   1639360   18283584   9%
/oradata
fas8060-nfs1:/vol_logs            9961472       128    9961344   1%
/logs
```

Erstellen Sie eine Vorlage für die Erstellung von Steuerdateien

Als nächstes müssen Sie eine controlfile-Vorlage erstellen. Der `backup controlfile to trace` Befehl erstellt Textbefehle, um eine Steuerdatei neu zu erstellen. Diese Funktion kann unter bestimmten Umständen hilfreich sein, um eine Datenbank aus einem Backup wiederherzustellen. Sie wird häufig bei Skripten verwendet, die Aufgaben wie das Klonen von Datenbanken ausführen.

1. Die Ausgabe des folgenden Befehls wird verwendet, um die Steuerdateien für die migrierte Datenbank neu zu erstellen.

```
SQL> alter database backup controlfile to trace as '/tmp/waffle.ctrl';
Database altered.
```

2. Nachdem die Steuerdateien erstellt wurden, kopieren Sie die Datei auf den neuen Server.

```
[oracle@jfsc3 tmp]$ scp oracle@jfsc1:/tmp/waffle.ctrl /tmp/
oracle@jfsc1's password:
waffle.ctrl                                               100% 5199
5.1KB/s   00:00
```

Parameterdatei sichern

In der neuen Umgebung ist auch eine Parameterdatei erforderlich. Die einfachste Methode ist, aus dem aktuellen spfile oder pfile ein pfile zu erstellen. In diesem Beispiel verwendet die Quelldatenbank eine spfile.

```
SQL> create pfile='/tmp/waffle.tmp.pfile' from spfile;
File created.
```

Oratab-Eintrag erstellen

Die Erstellung eines Oratab-Eintrags ist für das ordnungsgemäße Funktionieren von Dienstprogrammen wie oraenv erforderlich. Führen Sie den folgenden Schritt aus, um einen Oratab-Eintrag zu erstellen.

```
WAFFLE:/orabin/product/12.1.0/dbhome_1:N
```

Verzeichnisstruktur vorbereiten

Wenn die benötigten Verzeichnisse noch nicht vorhanden waren, müssen Sie sie erstellen, oder der Datenbankstartvorgang schlägt fehl. Um die Verzeichnisstruktur vorzubereiten, müssen Sie die folgenden Mindestanforderungen erfüllen.

```
[oracle@jfsc3 ~]$ . oraenv
ORACLE_SID = [oracle] ? WAFFLE
The Oracle base has been set to /orabin
[oracle@jfsc3 ~]$ cd $ORACLE_BASE
[oracle@jfsc3 orabin]$ cd admin
[oracle@jfsc3 admin]$ mkdir WAFFLE
[oracle@jfsc3 admin]$ cd WAFFLE
[oracle@jfsc3 WAFFLE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Aktualisierung der Parameterdatei

1. Um die Parameterdatei auf den neuen Server zu kopieren, führen Sie die folgenden Befehle aus. Der Standardspeicherort ist der `$ORACLE_HOME/dbs` Verzeichnis. In diesem Fall kann die pfile überall platziert werden. Sie wird nur als Zwischenschritt im Migrationsprozess genutzt.

```
[oracle@jfsc3 admin]$ scp oracle@jfsc1:/tmp/waffle.tmp.pfile
$ORACLE_HOME/dbs/waffle.tmp.pfile
oracle@jfsc1's password:
waffle.pfile                                          100%  916
0.9KB/s   00:00
```

1. Bearbeiten Sie die Datei nach Bedarf. Wenn sich beispielsweise der Speicherort des Archivprotokolls geändert hat, muss das pfile entsprechend dem neuen Speicherort geändert werden. In diesem Beispiel werden nur die Steuerdateien verschoben, zum Teil, um sie zwischen Protokoll- und Datendateisystemen zu verteilen.

```
[root@jfsc1 tmp]# cat waffle.pfile
WAFFLE.__data_transfer_cache_size=0
WAFFLE.__db_cache_size=507510784
WAFFLE.__java_pool_size=4194304
WAFFLE.__large_pool_size=20971520
WAFFLE.__oracle_base='/orabin'#ORACLE_BASE set from environment
WAFFLE.__pga_aggregate_target=268435456
WAFFLE.__sga_target=805306368
WAFFLE.__shared_io_pool_size=29360128
WAFFLE.__shared_pool_size=234881024
WAFFLE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/WAFFLE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/oradata//WAFFLE/control01.ctl','/oradata//WAFFLE/contr
ol02.ctl'
*.control_files='/oradata/WAFFLE/control01.ctl','/logs/WAFFLE/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='WAFFLE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=WAFFLEXDB)'
*.log_archive_dest_1='LOCATION=/logs/WAFFLE/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

2. Nachdem die Bearbeitungen abgeschlossen sind, erstellen Sie auf Basis dieses pfile ein spfile.

```
SQL> create spfile from pfile='waffle.tmp.pfile';
File created.
```

Erstellen Sie Steuerdateien neu

In einem vorherigen Schritt wird die Ausgabe von angezeigt `backup controlfile to trace` Wurde auf den neuen Server kopiert. Der spezifische Teil des erforderlichen Ausgangs ist der `controlfile recreation` Befehl. Diese Informationen finden Sie in der Datei unter dem markierten Abschnitt `Set #1. NORESETLOGS`. Es beginnt mit der Linie `create controlfile reuse database` Und sollte das Wort enthalten `noresetlogs`. Er endet mit dem Semikolon (;).

1. In diesem Beispiel liest die Datei wie folgt.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/logs/WAFFLE/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/logs/WAFFLE/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/logs/WAFFLE/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

2. Bearbeiten Sie dieses Skript wie gewünscht, um den neuen Speicherort der verschiedenen Dateien anzuzeigen. Beispielsweise können bestimmte Datendateien, von denen bekannt ist, dass sie eine hohe I/O-Last unterstützen, auf ein Filesystem auf einer hochperformanten Storage-Ebene umgeleitet werden. In anderen Fällen könnten die Änderungen lediglich aus Administratorgründen vorgenommen werden, wie z. B. die Isolierung der Datendateien einer bestimmten PDB in dedizierten Volumes.
3. In diesem Beispiel ist der `DATAFILE` Stanza bleibt unverändert, aber die Redo-Logs werden an einen neuen Speicherort in verschoben `/redo` Statt Speicherplatz für Archivprotokolle freizugeben `/logs`.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
SQL> CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
  2      MAXLOGFILES 16
  3      MAXLOGMEMBERS 3
  4      MAXDATAFILES 100
  5      MAXINSTANCES 8
  6      MAXLOGHISTORY 292
  7  LOGFILE
  8    GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  9    GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
 10    GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
 11  -- STANDBY LOGFILE
 12  DATAFILE
 13    '/oradata/WAFFLE/system01.dbf',
 14    '/oradata/WAFFLE/sysaux01.dbf',
 15    '/oradata/WAFFLE/undotbs01.dbf',
 16    '/oradata/WAFFLE/users01.dbf'
 17  CHARACTER SET WE8MSWIN1252
 18  ;
Control file created.
SQL>
```

Wenn Dateien falsch platziert oder Parameter falsch konfiguriert sind, werden Fehler generiert, die angeben, was repariert werden muss. Die Datenbank ist gemountet, aber noch nicht geöffnet und kann nicht geöffnet werden, da die verwendeten Datendateien noch als Hot Backup-Modus markiert sind. Um die Datenbankkonsistenz zu gewährleisten, müssen zunächst Archivprotokolle angewendet werden.

Erste Protokollreplizierung

Es ist mindestens ein Protokollantwort erforderlich, um die Datendateien konsistent zu gestalten. Es stehen zahlreiche Optionen zur Wiedergabe von Protokollen zur Verfügung. In einigen Fällen kann der ursprüngliche Speicherort des Archivprotokolls auf dem ursprünglichen Server über NFS freigegeben werden, und die Protokollantwort kann direkt erfolgen. In anderen Fällen müssen die Archivprotokolle kopiert werden.

Zum Beispiel, eine einfache `scp` Der Vorgang kann alle aktuellen Protokolle vom Quellserver auf den Migrationsserver kopieren:

```
[oracle@jfsc3 arch]$ scp jfsc1:/logs/WAFFLE/arch/* ./
oracle@jfsc1's password:
1_22_912662036.dbf                                          100%   47MB
47.0MB/s   00:01
1_23_912662036.dbf                                          100%   40MB
40.4MB/s   00:00
1_24_912662036.dbf                                          100%   45MB
45.4MB/s   00:00
1_25_912662036.dbf                                          100%   41MB
40.9MB/s   00:01
1_26_912662036.dbf                                          100%   39MB
39.4MB/s   00:00
1_27_912662036.dbf                                          100%   39MB
38.7MB/s   00:00
1_28_912662036.dbf                                          100%   40MB
40.1MB/s   00:01
1_29_912662036.dbf                                          100%   17MB
16.9MB/s   00:00
1_30_912662036.dbf                                          100%  636KB
636.0KB/s   00:00
```

Erste Protokollwiedergabe

Nachdem sich die Dateien im Archiv-Log-Speicherort befinden, können sie mit dem Befehl wiedergegeben werden `recover database until cancel` Gefolgt von der Antwort `AUTO` Um alle verfügbaren Protokolle automatisch wiederzugeben.

```
SQL> recover database until cancel;
ORA-00279: change 382713 generated at 05/24/2016 09:00:54 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_23_912662036.dbf
ORA-00280: change 382713 for thread 1 is in sequence #23
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 405712 generated at 05/24/2016 15:01:05 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_24_912662036.dbf
ORA-00280: change 405712 for thread 1 is in sequence #24
ORA-00278: log file '/logs/WAFFLE/arch/1_23_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
ORA-00278: log file '/logs/WAFFLE/arch/1_30_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_31_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

Die endgültige Antwort des Archivprotokolls meldet einen Fehler. Dies ist jedoch normal. Das Protokoll zeigt das an `sqlplus` Ich habe eine bestimmte Protokolldatei gesucht und sie nicht gefunden. Der Grund dafür ist höchstwahrscheinlich, dass die Protokolldatei noch nicht existiert.

Wenn die Quelldatenbank vor dem Kopieren von Archivprotokollen heruntergefahren werden kann, muss dieser Schritt nur einmal durchgeführt werden. Die Archivprotokolle werden kopiert und eingespielt. Anschließend kann der Prozess direkt zum Umstellungsprozess fortgesetzt werden, der die kritischen Wiederherstellungsprotokolle repliziert.

Inkrementelle Protokollreplikation und -Wiedergabe

In den meisten Fällen erfolgt die Migration nicht sofort. Es kann Tage oder sogar Wochen bis zum Abschluss des Migrationsprozesses dauern. Das bedeutet, dass die Protokolle kontinuierlich an die Replikatdatenbank gesendet und erneut eingespielt werden müssen. Bei Ankunft der Umstellung müssen daher nur minimale Daten übertragen und erneut eingespielt werden.

Dies kann auf viele Arten per Skript gesteuert werden, aber eine der beliebtesten Methoden ist die Verwendung von rsync, einem gemeinsamen Dateireplikationsdienstprogramm. Die sicherste Methode, dieses Dienstprogramm zu verwenden, ist es als Daemon zu konfigurieren. Beispiel: Der `rsyncd.conf` Die folgende Datei zeigt, wie eine Ressource mit dem Namen erstellt wird `waffle.arch` Der Zugriff erfolgt mit Oracle-Benutzeranmeldeinformationen und ist zugeordnet `/logs/WAFFLE/arch`. Am wichtigsten ist jedoch, dass die

Ressource schreibgeschützt ist, wodurch die Produktionsdaten gelesen, aber nicht verändert werden können.

```
[root@jfsc1 arch]# cat /etc/rsyncd.conf
[waffle.arch]
   uid=oracle
   gid=dba
   path=/logs/WAFFLE/arch
   read only = true
[root@jfsc1 arch]# rsync --daemon
```

Mit dem folgenden Befehl wird das Archivprotokollziel des neuen Servers mit der rsync-Ressource synchronisiert `waffle.arch` Auf dem ursprünglichen Server. Der `t` Argument in `rsync - potg` Führt dazu, dass die Dateiliste anhand des Zeitstempels verglichen wird und nur neue Dateien kopiert werden. Dieser Prozess bietet eine inkrementelle Aktualisierung des neuen Servers. Dieser Befehl kann auch in cron so geplant werden, dass er regelmäßig ausgeführt wird.

```
[oracle@jfsc3 arch]$ rsync -potg --stats --progress jfsc1::waffle.arch/*
/logs/WAFFLE/arch/
1_31_912662036.dbf
      650240 100%  124.02MB/s    0:00:00 (xfer#1, to-check=8/18)
1_32_912662036.dbf
     4873728 100%  110.67MB/s    0:00:00 (xfer#2, to-check=7/18)
1_33_912662036.dbf
     4088832 100%   50.64MB/s    0:00:00 (xfer#3, to-check=6/18)
1_34_912662036.dbf
     8196096 100%   54.66MB/s    0:00:00 (xfer#4, to-check=5/18)
1_35_912662036.dbf
    19376128 100%   57.75MB/s    0:00:00 (xfer#5, to-check=4/18)
1_36_912662036.dbf
       71680 100%  201.15kB/s    0:00:00 (xfer#6, to-check=3/18)
1_37_912662036.dbf
     1144320 100%    3.06MB/s    0:00:00 (xfer#7, to-check=2/18)
1_38_912662036.dbf
    35757568 100%   63.74MB/s    0:00:00 (xfer#8, to-check=1/18)
1_39_912662036.dbf
      984576 100%    1.63MB/s    0:00:00 (xfer#9, to-check=0/18)
Number of files: 18
Number of files transferred: 9
Total file size: 399653376 bytes
Total transferred file size: 75143168 bytes
Literal data: 75143168 bytes
Matched data: 0 bytes
File list size: 474
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 204
Total bytes received: 75153219
sent 204 bytes  received 75153219 bytes  150306846.00 bytes/sec
total size is 399653376  speedup is 5.32
```

Nachdem die Protokolle empfangen wurden, müssen sie erneut abgespielt werden. Frühere Beispiele zeigen die Verwendung von sqlplus zum manuellen Ausführen `recover database until cancel`Ein Prozess, der leicht automatisiert werden kann. Das hier abgebildete Beispiel verwendet das in beschriebene Skript "Protokolle in der Datenbank wiedergeben". Die Skripte akzeptieren ein Argument, das die Datenbank angibt, die einen Wiedergabevorgang erfordert. Damit kann dasselbe Skript bei einer Migration mit mehreren Datenbanken verwendet werden.

```
[oracle@jfsc3 logs]$ ./replay.logs.pl WAFFLE
ORACLE_SID = [WAFFLE] ? The Oracle base remains unchanged with value
/orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu May 26 10:47:16 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 814256 generated at 05/26/2016 04:52:30 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_32_912662036.dbf
ORA-00280: change 814256 for thread 1 is in sequence #32
ORA-00278: log file '/logs/WAFFLE/arch/1_31_912662036.dbf' no longer
needed for
this recovery
ORA-00279: change 814780 generated at 05/26/2016 04:53:04 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_33_912662036.dbf
ORA-00280: change 814780 for thread 1 is in sequence #33
ORA-00278: log file '/logs/WAFFLE/arch/1_32_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
ORA-00278: log file '/logs/WAFFLE/arch/1_39_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Umstellung

Wenn Sie bereit sind, in die neue Umgebung zu schneiden, müssen Sie eine abschließende Synchronisierung durchführen, die sowohl Archivprotokolle als auch Redo-Protokolle enthält. Wenn der ursprüngliche Speicherort des Wiederherstellungsprotokolls nicht bereits bekannt ist, kann er wie folgt identifiziert werden:

```
SQL> select member from v$logfile;
MEMBER
--------------------------------------------------------------------------------
------
/logs/WAFFLE/redo/redo01.log
/logs/WAFFLE/redo/redo02.log
/logs/WAFFLE/redo/redo03.log
```

1. Fahren Sie die Quelldatenbank herunter.
2. Führen Sie eine abschließende Synchronisierung der Archivprotokolle auf dem neuen Server mit der gewünschten Methode durch.
3. Die Wiederherstellungsprotokolle der Quelle müssen auf den neuen Server kopiert werden. In diesem Beispiel wurden die Wiederherstellungsprotokolle in ein neues Verzeichnis unter verschoben `/redo`.

   ```
   [oracle@jfsc3 logs]$ scp jfsc1:/logs/WAFFLE/redo/* /redo/
   oracle@jfsc1's password:
   redo01.log
   100%   50MB  50.0MB/s   00:01
   redo02.log
   100%   50MB  50.0MB/s   00:00
   redo03.log
   100%   50MB  50.0MB/s   00:00
   ```

4. In dieser Phase enthält die neue Datenbankumgebung alle Dateien, die als Quelle erforderlich sind. Die Archivprotokolle müssen ein letztes Mal wiedergegeben werden.

```
SQL> recover database until cancel;
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

5. Nach Abschluss müssen die Wiederherstellungsprotokolle erneut wiedergegeben werden. Wenn die Meldung angezeigt wird `Media recovery complete` Wird zurückgegeben, der Prozess ist erfolgreich und die Datenbanken sind synchronisiert und können geöffnet werden.

```
SQL> recover database;
Media recovery complete.
SQL> alter database open;
Database altered.
```

Protokollversand: ASM an Dateisystem

In diesem Beispiel wird die Verwendung von Oracle RMAN zur Migration einer Datenbank demonstriert. Es ähnelt dem vorherigen Beispiel des Dateisystems zum Protokollversand des Dateisystems, aber die Dateien auf ASM sind für den Host nicht sichtbar. Die einzigen Optionen für die Migration von Daten auf ASM-Geräten sind entweder die Verlagerung der ASM-LUN oder die Durchführung der Kopiervorgänge mithilfe von Oracle RMAN.

Auch wenn RMAN für das Kopieren von Dateien aus Oracle ASM erforderlich ist, ist die Verwendung von RMAN nicht auf ASM beschränkt. Mit RMAN können beliebige Storage-Typen zu beliebigen anderen Storage-Typen migriert werden.

Dieses Beispiel zeigt die Verlagerung einer Datenbank namens PANCAKE aus dem ASM-Speicher in ein normales Dateisystem, das sich auf einem anderen Server in Pfaden befindet `/oradata` Und `/logs`.

Erstellen Sie eine Datenbanksicherung

Im ersten Schritt wird ein Backup der Datenbank erstellt, die auf einen alternativen Server migriert werden soll. Da die Quelle Oracle ASM verwendet, muss RMAN verwendet werden. Ein einfaches RMAN-Backup kann wie folgt durchgeführt werden. Diese Methode erstellt ein getaggtes Backup, das später im Verfahren von RMAN

leicht identifiziert werden kann.

Der erste Befehl definiert den Zieltyp für das Backup und den zu verwendenden Speicherort. Die zweite initiiert nur die Sicherung der Datendateien.

```
RMAN> configure channel device type disk format '/rman/pancake/%U';
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters are successfully stored
RMAN> backup database tag 'ONTAP_MIGRATION';
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=251 device type=DISK
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/PANCAKE/system01.dbf
input datafile file number=00002 name=+ASM0/PANCAKE/sysaux01.dbf
input datafile file number=00003 name=+ASM0/PANCAKE/undotbs101.dbf
input datafile file number=00004 name=+ASM0/PANCAKE/users01.dbf
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1gr6c161_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:03
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1hr6c164_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Sicherungscontrolfile

Im weiteren Verlauf des Verfahrens wird eine Sicherungscontrolfile benötigt `duplicate database` Betrieb.

```
RMAN> backup current controlfile format '/rman/pancake/ctrl.bkp';
Starting backup at 24-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/ctrl.bkp tag=TAG20160524T032651 comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Parameterdatei sichern

In der neuen Umgebung ist auch eine Parameterdatei erforderlich. Die einfachste Methode ist, aus dem aktuellen spfile oder pfile ein pfile zu erstellen. In diesem Beispiel verwendet die Quelldatenbank eine spfile.

```
RMAN> create pfile='/rman/pancake/pfile' from spfile;
Statement processed
```

Skript zum Umbenennen der ASM-Datei

Mehrere aktuell in den Steuerdateien definierte Dateispeicherorte ändern sich, wenn die Datenbank verschoben wird. Mit dem folgenden Skript wird ein RMAN-Skript erstellt, um den Prozess zu vereinfachen. Dieses Beispiel zeigt eine Datenbank mit einer sehr kleinen Anzahl von Datendateien, aber in der Regel enthalten Datenbanken Hunderte oder gar Tausende von Datendateien.

Dieses Skript finden Sie in "Namenskonvertierung von ASM in Dateisystem" Und es tut zwei Dinge.

Zuerst erstellt es einen Parameter, um die Speicherort des Wiederherstellungsprotokolls neu zu definieren `log_file_name_convert`. Es handelt sich im Wesentlichen um eine Liste von abwechselnden Feldern. Das erste Feld ist der Speicherort eines aktuellen Wiederherstellungsprotokolls und das zweite Feld ist der Speicherort auf dem neuen Server. Das Muster wird dann wiederholt.

Die zweite Funktion ist die Bereitstellung einer Vorlage für die Umbenennung von Datendateien. Das Skript führt eine Schleife durch die Datendateien durch, ruft den Namen und die Dateinummer ab und formatiert sie als RMAN-Skript. Dann macht es das gleiche mit den temporären Dateien. Das Ergebnis ist ein einfaches rman-Skript, das nach Bedarf bearbeitet werden kann, um sicherzustellen, dass die Dateien an dem gewünschten Speicherort wiederhergestellt werden.

```
SQL> @/rman/mk.rename.scripts.sql
Parameters for log file conversion:
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/NEW_PATH/redo01.log','+ASM0/PANCAKE/redo02.log',
'/NEW_PATH/redo02.log','+ASM0/PANCAKE/redo03.log', '/NEW_PATH/redo03.log'
rman duplication script:
run
{
set newname for datafile 1 to '+ASM0/PANCAKE/system01.dbf';
set newname for datafile 2 to '+ASM0/PANCAKE/sysaux01.dbf';
set newname for datafile 3 to '+ASM0/PANCAKE/undotbs101.dbf';
set newname for datafile 4 to '+ASM0/PANCAKE/users01.dbf';
set newname for tempfile 1 to '+ASM0/PANCAKE/temp01.dbf';
duplicate target database for standby backup location INSERT_PATH_HERE;
}
PL/SQL procedure successfully completed.
```

Erfassen Sie die Ausgabe dieses Bildschirms. Der `log_file_name_convert` Der Parameter wird wie unten beschrieben in pfile platziert. Die RMAN-Datendatei umbenennen und das doppelte Skript müssen entsprechend bearbeitet werden, um die Datendateien an den gewünschten Speicherorten zu platzieren. In diesem Beispiel werden sie alle in platziert `/oradata/pancake`.

```
run
{
set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
set newname for datafile 4 to '/oradata/pancake/users.dbf';
set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
duplicate target database for standby backup location '/rman/pancake';
}
```

Verzeichnisstruktur vorbereiten

Die Skripte sind fast fertig zur Ausführung, aber zuerst muss die Verzeichnisstruktur vorhanden sein. Wenn die benötigten Verzeichnisse nicht bereits vorhanden sind, müssen sie erstellt werden, oder der Datenbankstartvorgang schlägt fehl. Das folgende Beispiel gibt die Mindestanforderungen wieder.

```
[oracle@jfsc2 ~]$ mkdir /oradata/pancake
[oracle@jfsc2 ~]$ mkdir /logs/pancake
[oracle@jfsc2 ~]$ cd /orabin/admin
[oracle@jfsc2 admin]$ mkdir PANCAKE
[oracle@jfsc2 admin]$ cd PANCAKE
[oracle@jfsc2 PANCAKE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Oratab-Eintrag erstellen

Der folgende Befehl ist für Dienstprogramme wie oraenv erforderlich, um ordnungsgemäß zu funktionieren.

```
PANCAKE:/orabin/product/12.1.0/dbhome_1:N
```

Parameteraktualisierungen

Die gespeicherte pfile muss aktualisiert werden, um alle Pfadänderungen auf dem neuen Server widerzuspiegeln. Die Änderungen des Datendateipfads werden durch das RMAN-Duplizierungsskript geändert, und fast alle Datenbanken erfordern Änderungen am `control_files` Und `log_archive_dest` Parameter. Es können auch Prüfdateipositionen vorhanden sein, die geändert werden müssen, und Parameter wie `db_create_file_dest` Ist außerhalb von ASM möglicherweise nicht relevant. Ein erfahrener DBA sollte die vorgeschlagenen Änderungen sorgfältig prüfen, bevor er fortfahren kann.

In diesem Beispiel sind die wichtigsten Änderungen die Speicherorte der Steuerdatei, das Protokollarchivziel und das Hinzufügen des `log_file_name_convert` Parameter.

```
PANCAKE.__data_transfer_cache_size=0
PANCAKE.__db_cache_size=545259520
PANCAKE.__java_pool_size=4194304
PANCAKE.__large_pool_size=25165824
PANCAKE.__oracle_base='/orabin'#ORACLE_BASE set from environment
PANCAKE.__pga_aggregate_target=268435456
PANCAKE.__sga_target=805306368
PANCAKE.__shared_io_pool_size=29360128
PANCAKE.__shared_pool_size=192937984
PANCAKE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/PANCAKE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='+ASM0/PANCAKE/control01.ctl','+ASM0/PANCAKE/control02.ctl
'
*.control_files='/oradata/pancake/control01.ctl','/logs/pancake/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='PANCAKE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=PANCAKEXDB)'
*.log_archive_dest_1='LOCATION=+ASM1'
*.log_archive_dest_1='LOCATION=/logs/pancake'
*.log_archive_format='%t_%s_%r.dbf'
'/logs/path/redo02.log'
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/logs/pancake/redo01.log', '+ASM0/PANCAKE/redo02.log',
'/logs/pancake/redo02.log', '+ASM0/PANCAKE/redo03.log',
'/logs/pancake/redo03.log'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Nachdem die neuen Parameter bestätigt wurden, müssen die Parameter wirksam werden. Es gibt mehrere Optionen, aber die meisten Kunden erstellen ein Spfile basierend auf dem Text pfile.

```
bash-4.1$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0 Production on Fri Jan 8 11:17:40 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> create spfile from pfile='/rman/pancake/pfile';
File created.
```

Startbezeichnung

Der letzte Schritt vor dem Replizieren der Datenbank ist, die Datenbankprozesse zu laden, aber nicht die Dateien zu mounten. In diesem Schritt können Probleme mit dem spfile offensichtlich werden. Wenn der `startup nomount` Befehl schlägt aufgrund eines Parameterfehlers fehl, es ist einfach herunterzufahren, die pfile-Vorlage zu korrigieren, sie als spfile neu zu laden und es erneut zu versuchen.

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             373296240 bytes
Database Buffers          423624704 bytes
Redo Buffers                5455872 bytes
```

Duplizieren Sie die Datenbank

Die Wiederherstellung des vorherigen RMAN-Backups am neuen Speicherort nimmt mehr Zeit in Anspruch als andere Schritte in diesem Prozess. Die Datenbank muss ohne Änderung der Datenbank-ID (DBID) oder Zurücksetzen der Protokolle dupliziert werden. Dadurch wird verhindert, dass Protokolle angewendet werden, was ein erforderlicher Schritt zur vollständigen Synchronisierung der Kopien ist.

Stellen Sie mit RMAN als AUX eine Verbindung zur Datenbank her, und geben Sie den Befehl Duplicate Database aus, indem Sie das in einem vorherigen Schritt erstellte Skript verwenden.

```
[oracle@jfsc2 pancake]$ rman auxiliary /
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 03:04:56
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to auxiliary database: PANCAKE (not mounted)
RMAN> run
2> {
3> set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
4> set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
5> set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
6> set newname for datafile 4 to '/oradata/pancake/users.dbf';
7> set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
```

```
8> duplicate target database for standby backup location '/rman/pancake';
9> }
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting Duplicate Db at 24-MAY-16
contents of Memory Script:
{
   restore clone standby controlfile from  '/rman/pancake/ctrl.bkp';
}
executing Memory Script
Starting restore at 24-MAY-16
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
channel ORA_AUX_DISK_1: restoring control file
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:01
output file name=/oradata/pancake/control01.ctl
output file name=/logs/pancake/control02.ctl
Finished restore at 24-MAY-16
contents of Memory Script:
{
   sql clone 'alter database mount standby database';
}
executing Memory Script
sql statement: alter database mount standby database
released channel: ORA_AUX_DISK_1
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
contents of Memory Script:
{
   set newname for tempfile  1 to
 "/oradata/pancake/temp.dbf";
   switch clone tempfile all;
   set newname for datafile  1 to
 "/oradata/pancake/pancake.dbf";
   set newname for datafile  2 to
 "/oradata/pancake/sysaux.dbf";
   set newname for datafile  3 to
 "/oradata/pancake/undotbs1.dbf";
   set newname for datafile  4 to
 "/oradata/pancake/users.dbf";
   restore
   clone database
   ;
```

```
}
executing Memory Script
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/pancake/temp.dbf in control file
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting restore at 24-MAY-16
using channel ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: starting datafile backup set restore
channel ORA_AUX_DISK_1: specifying datafile(s) to restore from backup set
channel ORA_AUX_DISK_1: restoring datafile 00001 to
/oradata/pancake/pancake.dbf
channel ORA_AUX_DISK_1: restoring datafile 00002 to
/oradata/pancake/sysaux.dbf
channel ORA_AUX_DISK_1: restoring datafile 00003 to
/oradata/pancake/undotbs1.dbf
channel ORA_AUX_DISK_1: restoring datafile 00004 to
/oradata/pancake/users.dbf
channel ORA_AUX_DISK_1: reading from backup piece
/rman/pancake/1gr6c161_1_1
channel ORA_AUX_DISK_1: piece handle=/rman/pancake/1gr6c161_1_1
tag=ONTAP_MIGRATION
channel ORA_AUX_DISK_1: restored backup piece 1
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:07
Finished restore at 24-MAY-16
contents of Memory Script:
{
   switch clone datafile all;
}
executing Memory Script
datafile 1 switched to datafile copy
input datafile copy RECID=5 STAMP=912655725 file
name=/oradata/pancake/pancake.dbf
datafile 2 switched to datafile copy
input datafile copy RECID=6 STAMP=912655725 file
name=/oradata/pancake/sysaux.dbf
datafile 3 switched to datafile copy
input datafile copy RECID=7 STAMP=912655725 file
name=/oradata/pancake/undotbs1.dbf
datafile 4 switched to datafile copy
input datafile copy RECID=8 STAMP=912655725 file
name=/oradata/pancake/users.dbf
Finished Duplicate Db at 24-MAY-16
```

Erste Protokollreplizierung

Sie müssen die Änderungen nun von der Quelldatenbank an einen neuen Speicherort senden. Dies kann eine Kombination von Schritten erfordern. Die einfachste Methode wäre, RMAN auf der Quelldatenbank zu haben, um Archivprotokolle auf eine freigegebene Netzwerkverbindung zu schreiben. Wenn ein freigegebener Speicherort nicht verfügbar ist, verwenden Sie RMAN zum Schreiben auf ein lokales Dateisystem und anschließend rcp oder rsync zum Kopieren der Dateien.

In diesem Beispiel ist der `/rman` Verzeichnis ist eine NFS-Freigabe, die sowohl für die ursprüngliche als auch für die migrierte Datenbank verfügbar ist.

Ein wichtiges Thema ist hier die `disk format` Klausel. Das Festplattenformat des Backups ist `%h_%e_%a.dbf`, Das bedeutet, dass Sie das Format der Thread-Nummer, Sequenznummer und Aktivierungs-ID für die Datenbank verwenden müssen. Obwohl die Buchstaben unterschiedlich sind, entspricht dies der `log_archive_format='%t_%s_%r.dbf` Parameter in pfile. Mit diesem Parameter werden auch Archivprotokolle im Format Thread-Nummer, Sequenznummer und Aktivierungs-ID angegeben. Das Endergebnis ist, dass die Protokolldatei-Backups auf der Quelle eine Benennungskonvention verwenden, die von der Datenbank erwartet wird. Dadurch werden z. B. Operationen wie die `recover database` Viel einfacher, weil sqlplus richtig vorwegnimmt die Namen der Archiv-Protokolle wiedergegeben werden.

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/arch/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
released channel: ORA_DISK_1
RMAN> backup as copy archivelog from time 'sysdate-2';
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=373 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=70 STAMP=912658508
output file name=/rman/pancake/logship/1_54_912576125.dbf RECID=123
STAMP=912659482
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=29 STAMP=912654101
output file name=/rman/pancake/logship/1_41_912576125.dbf RECID=124
STAMP=912659483
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=33 STAMP=912654688
output file name=/rman/pancake/logship/1_45_912576125.dbf RECID=152
STAMP=912659514
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=36 STAMP=912654809
output file name=/rman/pancake/logship/1_47_912576125.dbf RECID=153
STAMP=912659515
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Erste Protokollwiedergabe

Nachdem sich die Dateien im Archiv-Log-Speicherort befinden, können sie mit dem Befehl wiedergegeben werden `recover database until cancel` Gefolgt von der Antwort `AUTO` Um alle verfügbaren Protokolle automatisch wiederzugeben. Die Parameterdatei leitet derzeit Archivprotokolle an `/logs/archive`, Aber dies stimmt nicht mit dem Speicherort überein, an dem RMAN zum Speichern von Protokollen verwendet wurde. Der Speicherort kann vor der Wiederherstellung der Datenbank wie folgt vorübergehend umgeleitet werden.

```
SQL> alter system set log_archive_dest_1='LOCATION=/rman/pancake/logship'
scope=memory;
System altered.
SQL> recover standby database until cancel;
ORA-00279: change 560224 generated at 05/24/2016 03:25:53 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_49_912576125.dbf
ORA-00280: change 560224 for thread 1 is in sequence #49
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 560353 generated at 05/24/2016 03:29:17 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_50_912576125.dbf
ORA-00280: change 560353 for thread 1 is in sequence #50
ORA-00278: log file '/rman/pancake/logship/1_49_912576125.dbf' no longer
needed
for this recovery
...
ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
ORA-00278: log file '/rman/pancake/logship/1_53_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_54_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

Die endgültige Antwort des Archivprotokolls meldet einen Fehler. Dies ist jedoch normal. Der Fehler zeigt an, dass sqlplus eine bestimmte Protokolldatei gesucht und nicht gefunden hat. Der Grund dafür ist sehr wahrscheinlich, dass die Protokolldatei noch nicht existiert.

Wenn die Quelldatenbank vor dem Kopieren von Archivprotokollen heruntergefahren werden kann, muss dieser Schritt nur einmal durchgeführt werden. Die Archivprotokolle werden kopiert und eingespielt. Anschließend kann der Prozess direkt zum Umstellungsprozess fortgesetzt werden, der die kritischen Wiederherstellungsprotokolle repliziert.

Inkrementelle Protokollreplikation und -Wiedergabe

In den meisten Fällen erfolgt die Migration nicht sofort. Es kann Tage oder sogar Wochen bis zum Abschluss des Migrationsprozesses dauern. Das bedeutet, dass die Protokolle kontinuierlich an die Replikatdatenbank gesendet und wieder eingespielt werden müssen. So ist sichergestellt, dass bei der Umstellung nur minimale Daten übertragen und eingespielt werden müssen.

Dieser Prozess kann einfach per Skript ausgeführt werden. Beispielsweise kann der folgende Befehl für die

ursprüngliche Datenbank geplant werden, um sicherzustellen, dass der für den Protokollversand verwendete Speicherort fortlaufend aktualisiert wird.

```
[oracle@jfsc1 pancake]$ cat copylogs.rman
configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
backup as copy archivelog from time 'sysdate-2';
```

```
[oracle@jfsc1 pancake]$ rman target / cmdfile=copylogs.rman
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 04:36:19
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: PANCAKE (DBID=3574534589)
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=369 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:22
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=124 STAMP=912659483
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:23
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_41_912576125.dbf
continuing other job steps, job failed will not be re-run
```

```
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:55
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:57
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
Recovery Manager complete.
```

Nachdem die Protokolle empfangen wurden, müssen sie erneut abgespielt werden. Frühere Beispiele zeigten die Verwendung von sqlplus zum manuellen Ausführen `recover database until cancel`, Die leicht automatisiert werden kann. Das hier abgebildete Beispiel verwendet das in beschriebene Skript "Wiedergabe von Protokollen in der Standby-Datenbank". Das Skript akzeptiert ein Argument, das die Datenbank angibt, für die eine Wiedergabeoperation erforderlich ist. Bei diesem Prozess kann dasselbe Skript für eine Migration mit mehreren Datenbanken verwendet werden.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 04:47:10 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 562219 generated at 05/24/2016 04:15:08 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_55_912576125.dbf
ORA-00280: change 562219 for thread 1 is in sequence #55
ORA-00278: log file '/rman/pancake/logship/1_54_912576125.dbf' no longer
needed for this recovery
ORA-00279: change 562370 generated at 05/24/2016 04:19:18 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_56_912576125.dbf
ORA-00280: change 562370 for thread 1 is in sequence #56
ORA-00278: log file '/rman/pancake/logship/1_55_912576125.dbf' no longer
needed for this recovery
...
ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
ORA-00278: log file '/rman/pancake/logship/1_64_912576125.dbf' no longer
needed for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_65_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Umstellung

Wenn Sie bereit sind, in die neue Umgebung zu schneiden, müssen Sie eine abschließende Synchronisierung durchführen. Bei der Arbeit mit normalen Dateisystemen ist es leicht sicherzustellen, dass die migrierte Datenbank zu 100 % mit dem Original synchronisiert wird, da die ursprünglichen Wiederherstellungsprotokolle kopiert und wiedergegeben werden. Es gibt keinen guten Weg, dies mit ASM zu tun. Nur die Archivprotokolle können einfach wiederaufgenommen werden. Um sicherzustellen, dass keine Daten verloren gehen, muss das endgültige Herunterfahren der ursprünglichen Datenbank sorgfältig durchgeführt werden.

1. Zunächst muss die Datenbank stillgelegt werden, um sicherzustellen, dass keine Änderungen vorgenommen werden. Diese Stilllegung kann die Deaktivierung geplanter Vorgänge, das Herunterfahren von Listenern und/oder das Herunterfahren von Anwendungen umfassen.
2. Nach diesem Schritt erstellen die meisten DBAs eine Dummy-Tabelle, die als Marker für das Herunterfahren dient.
3. Erzwingen Sie eine Protokollarchivierung, um sicherzustellen, dass die Erstellung der Dummy-Tabelle in den Archivprotokollen aufgezeichnet wird. Führen Sie dazu die folgenden Befehle aus:

```
SQL> create table cutovercheck as select * from dba_users;
Table created.
SQL> alter system archive log current;
System altered.
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
```

4. Führen Sie die folgenden Befehle aus, um die letzten Archivprotokolle zu kopieren. Die Datenbank muss verfügbar, aber nicht geöffnet sein.

```
SQL> startup mount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
Database mounted.
```

5. Um die Archivprotokolle zu kopieren, führen Sie die folgenden Befehle aus:

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=8 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:24
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:58
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:59:00
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
```

6. Geben Sie abschließend die restlichen Archivprotokolle auf dem neuen Server wieder.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 05:00:53 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed
for thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 563629 generated at 05/24/2016 04:55:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_66_912576125.dbf
ORA-00280: change 563629 for thread 1 is in sequence #66
ORA-00278: log file '/rman/pancake/logship/1_65_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_66_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

7. In dieser Phase sollten Sie alle Daten replizieren. Die Datenbank kann von einer Standby-Datenbank in eine aktive Betriebsdatenbank konvertiert und dann geöffnet werden.

```
SQL> alter database activate standby database;
Database altered.
SQL> alter database open;
Database altered.
```

8. Bestätigen Sie das Vorhandensein der Dummy-Tabelle und legen Sie sie dann ab.

```
SQL> desc cutovercheck
 Name                                      Null?    Type
 ----------------------------------------- --------
---------------------------
 USERNAME                                  NOT NULL VARCHAR2(128)
 USER_ID                                   NOT NULL NUMBER
 PASSWORD                                           VARCHAR2(4000)
 ACCOUNT_STATUS                            NOT NULL VARCHAR2(32)
 LOCK_DATE                                          DATE
 EXPIRY_DATE                                        DATE
 DEFAULT_TABLESPACE                        NOT NULL VARCHAR2(30)
 TEMPORARY_TABLESPACE                      NOT NULL VARCHAR2(30)
 CREATED                                   NOT NULL DATE
 PROFILE                                   NOT NULL VARCHAR2(128)
 INITIAL_RSRC_CONSUMER_GROUP                        VARCHAR2(128)
 EXTERNAL_NAME                                      VARCHAR2(4000)
 PASSWORD_VERSIONS                                  VARCHAR2(12)
 EDITIONS_ENABLED                                   VARCHAR2(1)
 AUTHENTICATION_TYPE                                VARCHAR2(8)
 PROXY_ONLY_CONNECT                                 VARCHAR2(1)
 COMMON                                             VARCHAR2(3)
 LAST_LOGIN                                         TIMESTAMP(9) WITH
TIME ZONE
 ORACLE_MAINTAINED                                  VARCHAR2(1)
SQL> drop table cutovercheck;
Table dropped.
```

Unterbrechungsfreie Migration von Wiederherstellungsprotokollen

Es gibt Zeiten, in denen eine Datenbank insgesamt korrekt organisiert ist, mit Ausnahme der Wiederherstellungsprotokolle. Dies kann aus vielen Gründen geschehen, von denen die häufigste im Zusammenhang mit Snapshots steht. Produkte wie SnapManager für Oracle, SnapCenter und das Storage Management Framework NetApp Snap Creator ermöglichen eine nahezu sofortige Wiederherstellung einer Datenbank, jedoch nur, wenn Sie den Zustand der Daten-File-Volumes zurücksetzen. Wenn Redo-Logs Speicherplatz mit den Datendateien teilen, kann Reversion nicht sicher ausgeführt werden, da es zur Zerstörung der Redo-Protokolle führen würde, was wahrscheinlich Datenverlust bedeutet. Daher müssen die Redo-Logs verschoben werden.

Dieses Verfahren ist einfach und unterbrechungsfrei.

Aktuelle Konfiguration des Wiederherstellungsprotokolls

1. Ermitteln Sie die Anzahl der Redo-Log-Gruppen und deren jeweilige Gruppennummern.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
rows selected.
```

2. Geben Sie die Größe der Wiederherstellungsprotokolle ein.

```
SQL> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
--------------------------------------------------------------------------------
--------
1 524288000
2 524288000
3 524288000
```

Erstellen Sie neue Protokolle

1. Erstellen Sie für jedes Redo-Protokoll eine neue Gruppe mit einer passenden Größe und Anzahl von Mitgliedern.

```
SQL> alter database add logfile ('/newredo0/redo01a.log',
'/newredo1/redo01b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo02a.log',
'/newredo1/redo02b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo03a.log',
'/newredo1/redo03b.log') size 500M;
Database altered.
SQL>
```

2. Überprüfen Sie die neue Konfiguration.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
4 /newredo0/redo01a.log
4 /newredo1/redo01b.log
5 /newredo0/redo02a.log
5 /newredo1/redo02b.log
6 /newredo0/redo03a.log
6 /newredo1/redo03b.log
12 rows selected.
```

Alte Protokolle ablegen

1. Löschen Sie die alten Protokolle (Gruppen 1, 2 und 3).

```
SQL> alter database drop logfile group 1;
Database altered.
SQL> alter database drop logfile group 2;
Database altered.
SQL> alter database drop logfile group 3;
Database altered.
```

2. Wenn ein Fehler auftritt, der verhindert, dass Sie ein aktives Protokoll ablegen, erzwingen Sie einen Wechsel zum nächsten Protokoll, um die Sperre freizugeben und einen globalen Kontrollpunkt zu erzwingen. Siehe folgendes Beispiel für diesen Prozess. Der Versuch, die Logfile-Gruppe 2, die sich am alten Speicherort befand, zu löschen, wurde abgelehnt, da noch aktive Daten in dieser Logdatei vorhanden waren.

```
SQL> alter database drop logfile group 2;
alter database drop logfile group 2
*
ERROR at line 1:
ORA-01623: log 2 is current log for instance NTAP (thread 1) - cannot
drop
ORA-00312: online log 2 thread 1: '/redo0/NTAP/redo02a.log'
ORA-00312: online log 2 thread 1: '/redo1/NTAP/redo02b.log'
```

3. Eine Protokollarchivierung, gefolgt von einem Kontrollpunkt, ermöglicht es Ihnen, die Protokolldatei zu löschen.

```
SQL> alter system archive log current;
System altered.
SQL> alter system checkpoint;
System altered.
SQL> alter database drop logfile group 2;
Database altered.
```

4. Löschen Sie anschließend die Protokolle aus dem Dateisystem. Sie sollten diesen Vorgang mit äußerster Sorgfalt durchführen.

Host-Daten werden kopiert

Wie bei der Migration auf Datenbankebene bietet auch die Migration auf Hostebene einen vom Storage-Anbieter unabhängigen Ansatz.

Mit anderen Worten, manchmal "einfach die Dateien kopieren" ist die beste Option.

Obwohl dieser Low-Tech-Ansatz zu einfach erscheint, bietet er doch erhebliche Vorteile, da keine spezielle Software erforderlich ist und die Originaldaten während des Prozesses sicher unberührt bleiben. Die primäre Einschränkung besteht darin, dass eine Datenmigration auf Dateikopien einen unterbrechungsfreien Prozess darstellt, da die Datenbank vor Beginn des Kopiervorgangs heruntergefahren werden muss. Es gibt keine gute Möglichkeit, Änderungen innerhalb einer Datei zu synchronisieren, so dass die Dateien vollständig stillgelegt werden müssen, bevor das Kopieren beginnt.

Wenn das für einen Kopiervorgang erforderliche Herunterfahren nicht wünschenswert ist, ist die nächstbeste Host-basierte Option die Nutzung eines Logical Volume Managers (LVM). Es gibt viele LVM-Optionen, einschließlich Oracle ASM, alle mit ähnlichen Funktionen, aber auch mit einigen Einschränkungen, die berücksichtigt werden müssen. In den meisten Fällen lässt sich die Migration ohne Ausfallzeit und Unterbrechung durchführen.

Dateisystem wird kopiert

Der Nutzen einer einfachen Kopieroperation sollte nicht unterschätzt werden. Dieser Vorgang erfordert während des Kopierprozesses Ausfallzeiten, ist jedoch äußerst zuverlässig und erfordert keine besondere Expertise in Bezug auf Betriebssysteme, Datenbanken oder Speichersysteme. Darüber hinaus ist es sehr sicher, weil es die ursprünglichen Daten nicht beeinträchtigt. In der Regel ändert ein Systemadministrator die Quelldateisysteme, die schreibgeschützt gemountet werden, und startet dann einen Server neu, um zu gewährleisten, dass die aktuellen Daten nicht beschädigt werden können. Der Kopiervorgang kann mithilfe eines Skripts durchgeführt werden, um sicherzustellen, dass er so schnell wie möglich ohne das Risiko eines Benutzerfehlers ausgeführt wird. Da der I/O-Typ eine einfache, sequenzielle Datenübertragung ist, ist er eine äußerst effiziente Bandbreitennutzung.

Das folgende Beispiel zeigt eine Möglichkeit für eine sichere und schnelle Migration.

Umgebung

Die zu migrierende Umgebung ist wie folgt:

- Aktuelle Dateisysteme

```
ontap-nfs1:/host1_oradata       52428800  16196928  36231872  31%
/oradata
ontap-nfs1:/host1_logs          49807360    548032  49259328   2% /logs
```

- Neue Filesysteme

```
ontap-nfs1:/host1_logs_new      49807360       128  49807232   1%
/new/logs
ontap-nfs1:/host1_oradata_new   49807360       128  49807232   1%
/new/oradata
```

Überblick

Die Datenbank kann von einem Datenbankadministrator migriert werden, indem er die Datenbank herunterfährt und die Dateien kopiert. Der Prozess kann jedoch problemlos Skripte erstellen, wenn viele Datenbanken migriert werden müssen oder die Ausfallzeit minimiert werden muss. Die Verwendung von Skripten verringert zudem das Risiko von Benutzerfehlern.

Die Beispielskripte automatisieren die folgenden Vorgänge:

- Die Datenbank wird heruntergefahren
- Konvertieren der vorhandenen Dateisysteme in einen schreibgeschützten Zustand
- Kopieren aller Daten von der Quelle auf Zieldateisysteme, wobei alle Dateiberechtigungen erhalten bleiben
- Heben Sie das Mounten der alten und neuen Dateisysteme auf
- Erneutes Mounten der neuen Dateisysteme in denselben Pfaden wie die vorherigen Dateisysteme

Verfahren

1. Fahren Sie die Datenbank herunter.

```
[root@host1 current]# ./dbshut.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 15:58:48 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
```

2. Konvertieren Sie die Dateisysteme in schreibgeschützt. Dies ist mit einem Skript schneller möglich, wie in dargestellt "Dateisystem in schreibgeschützt konvertieren".

```
[root@host1 current]# ./mk.fs.readonly.pl /oradata
/oradata unmounted
/oradata mounted read-only
[root@host1 current]# ./mk.fs.readonly.pl /logs
/logs unmounted
/logs mounted read-only
```

3. Vergewissern Sie sich, dass die Dateisysteme jetzt schreibgeschützt sind.

```
ontap-nfs1:/host1_oradata on /oradata type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_logs on /logs type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

4. Synchronisieren Sie Dateisysteminhalte mit dem `rsync` Befehl.

```
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /oradata/ /new/oradata/
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
```

```
 10737426432 100%  153.50MB/s    0:01:06 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%   12.09MB/s    0:00:01 (xfer#2, to-check=9/13)
...
NTAP/undotbs02.dbf
  1073750016 100%  131.60MB/s    0:00:07 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    3.95MB/s    0:00:01 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  162204017.96 bytes/sec
total size is 18570092218  speedup is 1.00
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /logs/ /new/logs/
sending incremental file list
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%   95.98MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   49.45MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   44.68MB/s    0:00:01 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   68.03MB/s    0:00:00 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
```

```
sent 527098156 bytes  received 278 bytes  95836078.91 bytes/sec
total size is 527032832  speedup is 1.00
```

5. Heben Sie die Bereitstellung der alten Dateisysteme auf, und verschieben Sie die kopierten Daten. Dies ist mit einem Skript schneller möglich, wie in dargestellt "Ersetzen Sie Das Dateisystem".

```
[root@host1 current]# ./swap.fs.pl /logs,/new/logs
/new/logs unmounted
/logs unmounted
Updated /logs mounted
[root@host1 current]# ./swap.fs.pl /oradata,/new/oradata
/new/oradata unmounted
/oradata unmounted
Updated /oradata mounted
```

6. Vergewissern Sie sich, dass die neuen Dateisysteme in der Position sind.

```
ontap-nfs1:/host1_logs_new on /logs type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_oradata_new on /oradata type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

7. Starten Sie die Datenbank.

```
[root@host1 current]# ./dbstart.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 16:10:07 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
```

Vollständig automatisierte Umstellung

Dieses Beispielskript akzeptiert Argumente der Datenbank-SID gefolgt von gemeinsam getrennten Paaren von Dateisystemen. Für das oben abgebildete Beispiel wird der Befehl wie folgt ausgegeben:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
```

Wenn das Beispielskript ausgeführt wird, wird die folgende Sequenz ausgeführt. Er wird beendet, wenn in einem beliebigen Schritt ein Fehler auftritt:

1. Fahren Sie die Datenbank herunter.
2. Konvertieren Sie die aktuellen Dateisysteme in den schreibgeschützten Status.
3. Verwenden Sie jedes durch Kommas getrennte Paar von Dateisystemargumenten, und synchronisieren Sie das erste Dateisystem mit dem zweiten.
4. Entfernen Sie die früheren Dateisysteme.
5. Aktualisieren Sie die `/etc/fstab` Datei wie folgt:
   a. Erstellen Sie ein Backup bei `/etc/fstab.bak`.
   b. Kommentieren Sie die vorherigen Einträge für die vorherigen und neuen Dateisysteme.
   c. Erstellen Sie einen neuen Eintrag für das neue Dateisystem, das den alten Bereitstellungspunkt verwendet.
6. Mounten Sie die Dateisysteme.
7. Starten Sie die Datenbank.

Der folgende Text enthält ein Ausführungsbeispiel für dieses Skript:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:05:50 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
```

```
sending incremental file list
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%  185.40MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   81.34MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   70.42MB/s    0:00:00 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   47.08MB/s    0:00:01 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
sent 527098156 bytes  received 278 bytes  150599552.57 bytes/sec
total size is 527032832  speedup is 1.00
Succesfully replicated filesystem /logs to /new/logs
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
 10737426432 100%  176.55MB/s    0:00:58 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%    9.48MB/s    0:00:02 (xfer#2, to-check=9/13)
... NTAP/undotbs01.dbf
   309338112 100%   70.76MB/s    0:00:04 (xfer#9, to-check=2/13)
NTAP/undotbs02.dbf
  1073750016 100%  187.65MB/s    0:00:05 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    5.09MB/s    0:00:00 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
```

```
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  177725933.55 bytes/sec
total size is 18570092218  speedup is 1.00
Succesfully replicated filesystem /oradata to /new/oradata
swap 0 /logs /new/logs
/new/logs unmounted
/logs unmounted
Mounted updated /logs
Swapped filesystem /logs for /new/logs
swap 1 /oradata /new/oradata
/new/oradata unmounted
/oradata unmounted
Mounted updated /oradata
Swapped filesystem /oradata for /new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:08:59 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
[root@host1 current]#
```

Oracle ASM SPFile- und Passthwd-Migration

Eine Schwierigkeit beim Abschluss der ASM-Migration sind die ASM-spezifische SPFile- und die Passwort-Datei. Standardmäßig werden diese kritischen Metadatendateien auf der ersten definierten ASM-Laufwerksgruppe erstellt. Wenn eine bestimmte ASM-Datenträgergruppe evakuiert und entfernt werden muss, müssen die SPFile- und Passwortdatei, die diese ASM-Instanz regelt, verschoben werden.

Ein weiterer Anwendungsfall, in dem diese Dateien eventuell verschoben werden müssen, ist die Implementierung von Datenbankmanagement-Software wie beispielsweise SnapManager für Oracle oder dem SnapCenter Oracle Plug-in. Eine der Funktionen dieser Produkte besteht darin, eine Datenbank schnell wiederherzustellen, indem der Zustand der ASM-LUNs, die die Datendateien hosten, zurückgesetzt wird. Um dies zu tun, muss die ASM-Laufwerksgruppe offline geschaltet werden, bevor eine Wiederherstellung

durchgeführt werden kann. Dies ist kein Problem, solange die Datendateien einer Datenbank in einer dedizierten ASM-Datenträgergruppe isoliert sind.

Wenn diese Datenträgergruppe auch die ASM-Datei spfile/passwd enthält, kann die Datenträgergruppe nur offline geschaltet werden, wenn die gesamte ASM-Instanz heruntergefahren wird. Dies ist ein disruptiver Prozess, was bedeutet, dass die Datei spfile/passwd verschoben werden muss.

Umgebung

1. Datenbank-SID = TOAST
2. Aktuelle Datendateien auf `+DATA`
3. Aktuelle Logfiles und Controlfiles auf `+LOGS`
4. Neue ASM-Laufwerksgruppen als eingerichtet `+NEWDATA` Und `+NEWLOGS`

Speicherorte für ASM-SPfile/passwd-Dateien

Die Verlagerung dieser Dateien kann ohne Unterbrechungen erfolgen. Aus Sicherheitsgründen empfiehlt NetApp jedoch, die Datenbankumgebung herunterzufahren, damit Sie sicher sein können, dass die Dateien verschoben wurden und die Konfiguration ordnungsgemäß aktualisiert wird. Dieses Verfahren muss wiederholt werden, wenn mehrere ASM-Instanzen auf einem Server vorhanden sind.

Ermitteln Sie ASM-Instanzen

Ermitteln Sie die ASM-Instanzen anhand der in aufgezeichneten Daten `oratab` Datei: Die ASM-Instanzen werden durch ein +-Symbol gekennzeichnet.

```
-bash-4.1$ cat /etc/oratab | grep '^+'
+ASM:/orabin/grid:N             # line added by Agent
```

Auf diesem Server befindet sich eine ASM-Instanz namens +ASM.

Stellen Sie sicher, dass alle Datenbanken heruntergefahren werden

Der einzige sichtbare smon-Prozess sollte der sman für die verwendete ASM-Instanz sein. Ein weiterer smon-Prozess zeigt an, dass eine Datenbank noch läuft.

```
-bash-4.1$ ps -ef | grep smon
oracle     857     1  0 18:26 ?        00:00:00 asm_smon_+ASM
```

Der einzige smon-Prozess ist die ASM-Instanz selbst. Das bedeutet, dass keine anderen Datenbanken ausgeführt werden und ohne das Risiko einer Störung der Datenbankvorgänge sicher fortgesetzt werden kann.

Suchen Sie Dateien

Ermitteln Sie den aktuellen Speicherort der ASM-Datei und der Passwortdatei mithilfe des `spget` Und `pwget` Befehle.

```
bash-4.1$ asmcmd
ASMCMD> spget
+DATA/spfile.ora
```

```
ASMCMD> pwget --asm
+DATA/orapwasm
```

Beide Dateien befinden sich an der Basis des `+DATA` Festplattengruppe.

Dateien kopieren

Kopieren Sie die Dateien mit dem in die neue ASM-Datenträgergruppe `spcopy` Und `pwcopy` Befehle. Wenn die neue Laufwerksgruppe vor kurzem erstellt wurde und derzeit leer ist, muss sie möglicherweise zuerst gemountet werden.

```
ASMCMD> mount NEWDATA
```

```
ASMCMD> spcopy +DATA/spfile.ora +NEWDATA/spfile.ora
copying +DATA/spfile.ora -> +NEWDATA/spfilea.ora
```

```
ASMCMD> pwcopy +DATA/orapwasm +NEWDATA/orapwasm
copying +DATA/orapwasm -> +NEWDATA/orapwasm
```

Die Dateien wurden nun von kopiert `+DATA` Bis `+NEWDATA`.

ASM-Instanz aktualisieren

Die ASM-Instanz muss jetzt aktualisiert werden, um die Standortänderung widerzuspiegeln. Der `spset` Und `pwset` Befehle aktualisieren die zum Starten der ASM-Datenträgergruppe erforderlichen ASM-Metadaten.

```
ASMCMD> spset +NEWDATA/spfile.ora
ASMCMD> pwset --asm +NEWDATA/orapwasm
```

Aktivieren Sie ASM mit aktualisierten Dateien

Zu diesem Zeitpunkt verwendet die ASM-Instanz weiterhin die früheren Speicherorte dieser Dateien. Die Instanz muss neu gestartet werden, um ein erneutes Lesen der Dateien von ihren neuen Speicherorten zu erzwingen und Sperren für die vorherigen Dateien freizugeben.

```
-bash-4.1$ sqlplus / as sysasm
SQL> shutdown immediate;
ASM diskgroups volume disabled
ASM diskgroups dismounted
ASM instance shutdown
```

```
SQL> startup
ASM instance started
Total System Global Area 1140850688 bytes
Fixed Size                  2933400 bytes
Variable Size            1112751464 bytes
ASM Cache                  25165824 bytes
ORA-15032: not all alterations performed
ORA-15017: diskgroup "NEWDATA" cannot be mounted
ORA-15013: diskgroup "NEWDATA" is already mounted
```

Entfernen Sie alte spfile- und Passwortdateien

Wenn der Vorgang erfolgreich durchgeführt wurde, sind die vorherigen Dateien nicht mehr gesperrt und können jetzt entfernt werden.

```
-bash-4.1$ asmcmd
ASMCMD> rm +DATA/spfile.ora
ASMCMD> rm +DATA/orapwasm
```

Kopie von Oracle ASM zu ASM

Oracle ASM ist im Grunde ein schlankes kombiniertes Volume-Manager- und Dateisystem. Da das Dateisystem nicht sofort sichtbar ist, muss RMAN für Kopiervorgänge verwendet werden. Ein auf Kopien basierender Migrationsprozess ist zwar sicher und einfach, kann jedoch mit Unterbrechungen verbunden sein. Die Unterbrechung kann minimiert, aber nicht vollständig beseitigt werden.

Wenn Sie eine unterbrechungsfreie Migration einer ASM-basierten Datenbank wünschen, empfiehlt es sich, die ASM-Fähigkeit zu nutzen, um ASM-Extents auf neue LUNs auszugleichen, während die alten LUNs gelöscht werden. Dies ist im Allgemeinen sicher und unterbrechungsfrei, bietet aber keinen Ausweg. Wenn Funktions- oder Leistungsprobleme auftreten, besteht die einzige Möglichkeit darin, die Daten zurück zur Quelle zu migrieren.

Dieses Risiko kann vermieden werden, indem die Datenbank an den neuen Speicherort kopiert wird, anstatt Daten zu verschieben, sodass die Originaldaten nicht geändert werden. Die Datenbank kann vor der Inbetriebnahme vollständig an ihrem neuen Standort getestet werden, und die ursprüngliche Datenbank steht als Fallback-Option zur Verfügung, wenn Probleme gefunden werden.

Dieses Verfahren ist eine von vielen Optionen, die RMAN einbeziehen. Er ermöglicht einen zweistufigen Prozess, bei dem das erste Backup erstellt und später durch die Protokollwiedergabe synchronisiert wird. Dieser Prozess sollte die Downtime minimieren, da die Datenbank betriebsbereit bleibt und während der ersten Basiskopie Daten bereitgestellt werden können.

Datenbank kopieren

Oracle RMAN erstellt eine vollständige Kopie der Quelldatenbank der Ebene 0, die sich derzeit in der ASM-Datenträgergruppe befindet `+DATA` An den neuen Standort am `+NEWDATA`.

```
-bash-4.1$ rman target /
Recovery Manager: Release 12.1.0.2.0 - Production on Sun Dec 6 17:40:03
2015
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: TOAST (DBID=2084313411)
RMAN> backup as copy incremental level 0 database format '+NEWDATA' tag
'ONTAP_MIGRATION';
Starting backup at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=302 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
...
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
output file name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
tag=ONTAP_MIGRATION RECID=5 STAMP=897759622
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWDATA/TOAST/BACKUPSET/2015_12_06/nnsnn0_ontap_migration_0.262.89
7759623 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Schalter für Archivprotokoll erzwingen

Sie müssen einen Schalter für das Archivprotokoll erzwingen, um sicherzustellen, dass die Archivprotokolle alle Daten enthalten, die erforderlich sind, um die Kopie vollständig konsistent zu machen. Ohne diesen Befehl können Schlüsseldaten in den Wiederherstellungsprotokollen weiterhin vorhanden sein.

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Quelldatenbank herunterfahren

Die Unterbrechung beginnt in diesem Schritt, weil die Datenbank heruntergefahren und in einen schreibgeschützten Modus mit eingeschränktem Zugriff versetzt wird. Um die Quelldatenbank herunterzufahren, führen Sie die folgenden Befehle aus:

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                390073456 bytes
Database Buffers             406847488 bytes
Redo Buffers                   5455872 bytes
```

Backup von Controlfile

Sie müssen die controlfile sichern, falls Sie die Migration abbrechen und zum ursprünglichen Speicherort zurückkehren müssen. Eine Kopie der Backup-Steuerdatei ist nicht 100% erforderlich, aber es macht den Prozess des Rücksetzens der Datenbank-Speicherorte zurück an den ursprünglichen Speicherort einfacher.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=358 device type=DISK
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151206T174753 RECID=6
STAMP=897760073
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Parameteraktualisierungen

Der aktuelle spfile enthält Verweise auf die Steuerdateien an ihren aktuellen Speicherorten innerhalb der alten ASM-Datenträgergruppe. Es muss bearbeitet werden, was leicht durch das Bearbeiten einer Zwischenversion von pfile erfolgt.

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Aktualisieren Sie pfile

Aktualisieren Sie alle Parameter, die sich auf alte ASM-Datenträgergruppen beziehen, um die neuen Namen der ASM-Datenträgergruppen wiederzugeben. Speichern Sie dann die aktualisierte Datei pfile. Stellen Sie sicher, dass die `db_create` Parameter sind vorhanden.

Im folgenden Beispiel werden die Verweise auf angezeigt `+DATA` Die in geändert wurden `+NEWDATA` Sind gelb markiert. Zwei wichtige Parameter sind die `db_create` Parameter, die neue Dateien am richtigen Speicherort erstellen.

```
*.compatible='12.1.0.2.0'
*.control_files='+NEWLOGS/TOAST/CONTROLFILE/current.258.897683139'
*.db_block_size=8192
*. db_create_file_dest='+NEWDATA'
*. db_create_online_log_dest_1='+NEWLOGS'
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=+NEWLOGS'
*.log_archive_format='%t_%s_%r.dbf'
```

Init.ora-Datei aktualisieren

Die meisten ASM-basierten Datenbanken verwenden einen `init.ora` Datei befindet sich im `$ORACLE_HOME/dbs` Verzeichnis, das einen Punkt auf das Spfile auf der ASM-Datenträgergruppe darstellt. Diese Datei muss an einen Speicherort auf der neuen ASM-Datenträgergruppe umgeleitet werden.

```
-bash-4.1$ cd $ORACLE_HOME/dbs
-bash-4.1$ cat initTOAST.ora
SPFILE='+DATA/TOAST/spfileTOAST.ora'
```

Ändern Sie diese Datei wie folgt:

```
SPFILE=+NEWLOGS/TOAST/spfileTOAST.ora
```

Wiederherstellung der Parameterdatei

Der spfile kann nun mit den Daten in der bearbeiteten pfile gefüllt werden.

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Starten Sie die Datenbank, um neue spfile zu verwenden

Starten Sie die Datenbank, um sicherzustellen, dass sie jetzt den neu erstellten spfile verwendet und dass alle weiteren Änderungen an den Systemparametern korrekt aufgezeichnet werden.

```
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                373296240 bytes
Database Buffers             423624704 bytes
Redo Buffers                   5455872 bytes
```

Kontrolldatei wiederherstellen

Die von RMAN erstellte Backup-Controldatei kann auch direkt an dem im neuen spfile angegebenen Speicherort wiederhergestellt werden.

```
RMAN> restore controlfile from
'+DATA/TOAST/CONTROLFILE/current.258.897683139';
Starting restore at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=417 device type=DISK
channel ORA_DISK_1: copied control file copy
output file name=+NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
Finished restore at 06-DEC-15
```

Mounten Sie die Datenbank und überprüfen Sie die Verwendung der neuen Steuerdatei.

```
RMAN> alter database mount;
using target database control file instead of recovery catalog
Statement processed
```

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
+NEWLOGS/TOAST/CONTROLFILE/cur
                                                 rent.273.897761061
```

Protokollwiedergabe

Die Datenbank verwendet derzeit die Datendateien am alten Speicherort. Bevor die Kopie verwendet werden kann, müssen sie synchronisiert werden. Die Zeit während des ersten Kopiervorgangs ist verstrichen, und die Änderungen wurden hauptsächlich in den Archivprotokollen protokolliert. Diese Änderungen werden wie folgt repliziert:

1. Führen Sie ein inkrementelles RMAN-Backup durch, das die Archivprotokolle enthält.

```
RMAN> backup incremental level 1 format '+NEWLOGS' for recover of copy
with tag 'ONTAP_MIGRATION' database;
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=62 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
input datafile file number=00002
name=+DATA/TOAST/DATAFILE/sysaux.260.897683143
input datafile file number=00003
name=+DATA/TOAST/DATAFILE/undotbs1.257.897683145
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/ncsnn1_ontap_migration_0.267.
897762697 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

2. Wiederholen Sie das Protokoll.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 06-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001
name=+NEWDATA/TOAST/DATAFILE/system.259.897759609
recovering datafile copy file number=00002
name=+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
recovering datafile copy file number=00003
name=+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
recovering datafile copy file number=00004
name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
channel ORA_DISK_1: reading from backup piece
+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.8977626
93
channel ORA_DISK_1: piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Aktivierung

Die wiederhergestellte Steuerdatei verweist weiterhin auf die Datendateien am ursprünglichen Speicherort und enthält auch die Pfadinformationen für die kopierten Datendateien.

1. Um die aktiven Datendateien zu ändern, führen Sie den aus `switch database to copy` Befehl.

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/system.259.897759609"
datafile 2 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615"
datafile 3 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619"
datafile 4 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/users.258.897759623"
```

   Die aktiven Datendateien sind nun die kopierten Datendateien, aber es können immer noch Änderungen in den letzten Redo-Protokollen enthalten sein.

2. Um alle verbleibenden Protokolle wiederzugeben, führen Sie den aus `recover database` Befehl. Wenn die Meldung angezeigt wird `media recovery complete` Wird angezeigt, der Prozess war erfolgreich.

```
RMAN> recover database;
Starting recover at 06-DEC-15
using channel ORA_DISK_1
starting media recovery
media recovery complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Bei diesem Vorgang wurde nur der Speicherort der normalen Datendateien geändert. Die temporären Datendateien müssen umbenannt werden, müssen aber nicht kopiert werden, da sie nur temporär sind. Die Datenbank ist derzeit nicht verfügbar, sodass es keine aktiven Daten in den temporären Datendateien gibt.

3. Um die temporären Datendateien zu verschieben, geben Sie zuerst ihren Speicherort an.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
--------------------------------------------------------------------------------
--------
1 +DATA/TOAST/TEMPFILE/temp.263.897683145
```

4. Verschieben Sie temporäre Datendateien mithilfe eines RMAN-Befehls, der den neuen Namen für jede Datendatei festlegt. Bei Oracle Managed Files (OMF) ist der vollständige Name nicht erforderlich; die ASM-Datenträgergruppe reicht aus. Wenn die Datenbank geöffnet wird, verknüpft OMF mit dem entsprechenden Speicherort in der ASM-Datenträgergruppe. Um Dateien zu verschieben, führen Sie die folgenden Befehle aus:

```
run {
set newname for tempfile 1 to '+NEWDATA';
switch tempfile all;
}
```

```
RMAN> run {
2> set newname for tempfile 1 to '+NEWDATA';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to +NEWDATA in control file
```

Migration des Wiederherstellungsprotokolls

Der Migrationsprozess ist fast abgeschlossen, aber die Wiederherstellungsprotokolle befinden sich immer noch in der ursprünglichen ASM-Laufwerksgruppe. Wiederherstellungsprotokolle können nicht direkt verschoben werden. Stattdessen wird ein neuer Satz von Wiederherstellungsprotokollen erstellt und der Konfiguration hinzugefügt, gefolgt von einem Drop der alten Protokolle.

1. Ermitteln Sie die Anzahl der Redo-Log-Gruppen und deren jeweilige Gruppennummern.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
------------------------------------------------------------------------------
--------
1 +DATA/TOAST/ONLINELOG/group_1.261.897683139
2 +DATA/TOAST/ONLINELOG/group_2.259.897683139
3 +DATA/TOAST/ONLINELOG/group_3.256.897683139
```

2. Geben Sie die Größe der Wiederherstellungsprotokolle ein.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
------------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Erstellen Sie für jedes Redo-Protokoll eine neue Gruppe mit einer passenden Konfiguration. Wenn Sie OMF nicht verwenden, müssen Sie den vollständigen Pfad angeben. Dies ist auch ein Beispiel, das den verwendet `db_create_online_log` Parameter. Wie bereits gezeigt, wurde dieser Parameter auf +NEWLOGS gesetzt. Mit dieser Konfiguration können Sie die folgenden Befehle verwenden, um neue Online-Protokolle zu erstellen, ohne einen Dateispeicherort oder sogar eine bestimmte ASM-Datenträgergruppe angeben zu müssen.

```
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
```

4. Öffnen Sie die Datenbank.

```
SQL> alter database open;
Database altered.
```

5. Die alten Protokolle ablegen.

```
RMAN> alter database drop logfile group 1;
Statement processed
```

6. Wenn ein Fehler auftritt, der verhindert, dass Sie ein aktives Protokoll ablegen, erzwingen Sie einen Wechsel zum nächsten Protokoll, um die Sperre freizugeben und einen globalen Kontrollpunkt zu erzwingen. Ein Beispiel ist unten dargestellt. Der Versuch, die Logfile-Gruppe 3, die sich am alten Speicherort befand, zu löschen, wurde abgelehnt, da noch aktive Daten in dieser Logdatei vorhanden waren. Eine Protokollarchivierung nach einem Kontrollpunkt ermöglicht das Löschen der Protokolldatei.

```
RMAN> alter database drop logfile group 3;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 3 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 3 thread 1:
'+LOGS/TOAST/ONLINELOG/group_3.259.897563549'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 3;
Statement processed
```

7. Überprüfen Sie die Umgebung, um sicherzustellen, dass alle standortbasierten Parameter aktualisiert werden.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

8. Im folgenden Skript wird erläutert, wie dieser Prozess vereinfacht werden kann:

```
[root@host1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
+NEWDATA/TOAST/DATAFILE/system.259.897759609
+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
+NEWDATA/TOAST/DATAFILE/users.258.897759623
TOAST redo logs:
+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123
+NEWLOGS/TOAST/ONLINELOG/group_5.265.897763125
+NEWLOGS/TOAST/ONLINELOG/group_6.264.897763125
TOAST temp datafiles:
+NEWDATA/TOAST/TEMPFILE/temp.260.897763165
TOAST spfile
spfile                                 string
+NEWDATA/spfiletoast.ora
TOAST key parameters
control_files +NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
log_archive_dest_1 LOCATION=+NEWLOGS
db_create_file_dest +NEWDATA
db_create_online_log_dest_1 +NEWLOGS
```

9. Wenn die ASM-Datenträgergruppen vollständig evakuiert wurden, können sie jetzt mit abgehängt werden `asmcmd`. In vielen Fällen sind jedoch die Dateien, die zu anderen Datenbanken oder der ASM-Datei spfile/passwd gehören, noch vorhanden.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Kopie von Oracle ASM auf das Dateisystem

Das Verfahren zum Kopieren von Oracle ASM in ein Dateisystem ähnelt dem Verfahren zum Kopieren von ASM zu ASM mit ähnlichen Vorteilen und Einschränkungen. Der Hauptunterschied ist die Syntax der verschiedenen Befehle und Konfigurationsparameter bei der Verwendung eines sichtbaren Dateisystems im Gegensatz zu einer ASM-Datenträgergruppe.

Datenbank kopieren

Oracle RMAN wird verwendet, um eine (vollständige) Kopie der Quelldatenbank zu erstellen, die sich derzeit in der ASM-Datenträgergruppe befindet `+DATA` An den neuen Standort am `/oradata`.

```
RMAN> backup as copy incremental level 0 database format
'/oradata/TOAST/%U' tag 'ONTAP_MIGRATION';
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=377 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-
1_01r5fhjg tag=ONTAP_MIGRATION RECID=1 STAMP=911722099
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-
2_02r5fhjo tag=ONTAP_MIGRATION RECID=2 STAMP=911722106
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt tag=ONTAP_MIGRATION RECID=3 STAMP=911722113
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/oradata/TOAST/cf_D-TOAST_id-2098173325_04r5fhk5
tag=ONTAP_MIGRATION RECID=4 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting datafile copy
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-
4_05r5fhk6 tag=ONTAP_MIGRATION RECID=5 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/oradata/TOAST/06r5fhk7_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
```

Schalter für Archivprotokoll erzwingen

Der Wechsel des Archivprotokolls muss erzwungen werden, um sicherzustellen, dass die Archivprotokolle alle erforderlichen Daten enthalten, damit die Kopie vollständig konsistent ist. Ohne diesen Befehl können Schlüsseldaten in den Wiederherstellungsprotokollen weiterhin vorhanden sein. Um einen Archivprotokollschalter zu erzwingen, führen Sie den folgenden Befehl aus:

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Quelldatenbank herunterfahren

Die Unterbrechung beginnt in diesem Schritt, weil die Datenbank heruntergefahren und in einen schreibgeschützten Modus mit eingeschränktem Zugriff versetzt wird. Um die Quelldatenbank herunterzufahren, führen Sie die folgenden Befehle aus:

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Backup von Controlfile

Sichern Sie controlfiles, falls Sie die Migration abbrechen und zum ursprünglichen Speicherort zurückkehren müssen. Eine Kopie der Backup-Steuerdatei ist nicht 100% erforderlich, aber es macht den Prozess des Rücksetzens der Datenbank-Speicherorte zurück an den ursprünglichen Speicherort einfacher.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 08-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151208T194540 RECID=30
STAMP=897939940
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 08-DEC-15
```

Parameteraktualisierungen

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Aktualisieren Sie pfile

Alle Parameter, die sich auf alte ASM-Datenträgergruppen beziehen, sollten aktualisiert und in einigen Fällen gelöscht werden, wenn sie nicht mehr relevant sind. Aktualisieren Sie sie, um die neuen Dateisystempfade wiederzugeben, und speichern Sie die aktualisierte Datei pfile. Stellen Sie sicher, dass der vollständige Zielpfad aufgeführt ist. Um diese Parameter zu aktualisieren, führen Sie die folgenden Befehle aus:

```
*.audit_file_dest='/orabin/admin/TOAST/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/logs/TOAST/arch/control01.ctl','/logs/TOAST/redo/control
02.ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=/logs/TOAST/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Deaktivieren Sie die ursprüngliche init.ora-Datei

Diese Datei befindet sich im `$ORACLE_HOME/dbs` Verzeichnis und befindet sich in der Regel in einem pfile, das als Zeiger auf den spfile auf der ASM-Datenträgergruppe dient. Um sicherzustellen, dass der ursprüngliche Spfile nicht mehr verwendet wird, benennen Sie ihn um. Löschen Sie sie jedoch nicht, da diese Datei erforderlich ist, wenn die Migration abgebrochen werden muss.

```
[oracle@jfsc1 ~]$ cd $ORACLE_HOME/dbs
[oracle@jfsc1 dbs]$ cat initTOAST.ora
SPFILE='+ASM0/TOAST/spfileTOAST.ora'
[oracle@jfsc1 dbs]$ mv initTOAST.ora initTOAST.ora.prev
[oracle@jfsc1 dbs]$
```

Wiederherstellung der Parameterdatei

Dies ist der letzte Schritt bei der Verlagerung von Spfile. Der ursprüngliche spfile wird nicht mehr verwendet und die Datenbank wird derzeit mit der Zwischendatei gestartet (aber nicht gemountet). Der Inhalt dieser Datei kann wie folgt an den neuen Speicherort spfile geschrieben werden:

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Starten Sie die Datenbank, um neue spfile zu verwenden

Sie müssen die Datenbank starten, um die Sperren der Zwischendatei freizugeben und die Datenbank nur mit der neuen Datei spfile zu starten. Das Starten der Datenbank beweist auch, dass der neue spfile-Speicherort korrekt ist und seine Daten gültig sind.

```
RMAN> shutdown immediate;
Oracle instance shut down
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Kontrolldatei wiederherstellen

Auf dem Pfad wurde eine Sicherungscontroldatei erstellt `/tmp/TOAST.ctrl` Früher im Verfahren. Der neue spfile definiert die Speicherorte der controlfile als /`logfs/TOAST/ctrl/ctrlfile1.ctrl` Und `/logfs/TOAST/redo/ctrlfile2.ctrl`. Diese Dateien sind jedoch noch nicht vorhanden.

1. Mit diesem Befehl werden die controlfile-Daten auf den im spfile definierten Pfaden wiederhergestellt.

   ```
   RMAN> restore controlfile from '/tmp/TOAST.ctrl';
   Starting restore at 13-MAY-16
   using channel ORA_DISK_1
   channel ORA_DISK_1: copied control file copy
   output file name=/logs/TOAST/arch/control01.ctl
   output file name=/logs/TOAST/redo/control02.ctl
   Finished restore at 13-MAY-16
   ```

2. Geben Sie den Mount-Befehl ein, damit die Steuerdateien korrekt erkannt werden und gültige Daten enthalten.

```
RMAN> alter database mount;
Statement processed
released channel: ORA_DISK_1
```

Um den zu validieren `control_files` Parameter, führen Sie den folgenden Befehl aus:

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
/logs/TOAST/arch/control01.ctl
                                                 ,
/logs/TOAST/redo/control02.c
                                                 tl
```

Protokollwiedergabe

Die Datenbank verwendet derzeit die Datendateien am alten Speicherort. Bevor die Kopie verwendet werden kann, müssen die Datendateien synchronisiert werden. Die Zeit während des ersten Kopiervorgangs ist verstrichen, und Änderungen wurden hauptsächlich in den Archivprotokollen protokolliert. Diese Änderungen werden in den folgenden beiden Schritten repliziert.

1. Führen Sie ein inkrementelles RMAN-Backup durch, das die Archivprotokolle enthält.

```
RMAN>  backup incremental level 1 format '/logs/TOAST/arch/%U' for
recover of copy with tag 'ONTAP_MIGRATION' database;
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=124 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/logs/TOAST/arch/09r5fj8i_1_1 tag=ONTAP_MIGRATION
comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

2. Wiederholen Sie die Protokolle.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 13-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
recovering datafile copy file number=00002 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
recovering datafile copy file number=00003 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
recovering datafile copy file number=00004 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
channel ORA_DISK_1: reading from backup piece
/logs/TOAST/arch/09r5fj8i_1_1
channel ORA_DISK_1: piece handle=/logs/TOAST/arch/09r5fj8i_1_1
tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

Aktivierung

Die wiederhergestellte Steuerdatei verweist weiterhin auf die Datendateien am ursprünglichen Speicherort und enthält auch die Pfadinformationen für die kopierten Datendateien.

1. Um die aktiven Datendateien zu ändern, führen Sie den aus `switch database to copy` Befehl:

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSTEM_FNO-1_01r5fhjg"
datafile 2 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSAUX_FNO-2_02r5fhjo"
datafile 3 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt"
datafile 4 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-USERS_FNO-4_05r5fhk6"
```

2. Obwohl die Datendateien vollständig konsistent sein sollten, ist ein letzter Schritt erforderlich, um die verbleibenden Änderungen, die in den Online-Wiederherstellungsprotokollen aufgezeichnet werden, wiederzugeben. Verwenden Sie die `recover database` Befehl, um diese Änderungen erneut einzuspielen und die Kopie 100 % mit dem Original zu identisch zu machen. Die Kopie ist jedoch noch nicht geöffnet.

```
RMAN> recover database;
Starting recover at 13-MAY-16
using channel ORA_DISK_1
starting media recovery
archived log for thread 1 with sequence 28 is already on disk as file
+ASM0/TOAST/redo01.log
archived log file name=+ASM0/TOAST/redo01.log thread=1 sequence=28
media recovery complete, elapsed time: 00:00:00
Finished recover at 13-MAY-16
```

Temporäre Datendateien Verschieben

1. Ermitteln Sie den Speicherort der temporären Datendateien, die noch auf der ursprünglichen Laufwerksgruppe verwendet werden.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
------------------------------------------------------------------------
--------
1 +ASM0/TOAST/temp01.dbf
```

2. Um die Datendateien zu verschieben, führen Sie die folgenden Befehle aus. Wenn es viele Tempfiles gibt, verwenden Sie einen Texteditor, um den RMAN-Befehl zu erstellen, und schneiden Sie ihn dann aus und fügen Sie ihn ein.

```
RMAN> run {
2> set newname for tempfile 1 to '/oradata/TOAST/temp01.dbf';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/TOAST/temp01.dbf in control file
```

Migration des Wiederherstellungsprotokolls

Der Migrationsprozess ist fast abgeschlossen, aber die Wiederherstellungsprotokolle befinden sich immer noch in der ursprünglichen ASM-Laufwerksgruppe. Wiederherstellungsprotokolle können nicht direkt verschoben werden. Stattdessen wird ein neuer Satz von Wiederherstellungsprotokollen erstellt und der Konfiguration hinzugefügt, gefolgt von einem Drop der alten Protokolle.

1. Ermitteln Sie die Anzahl der Redo-Log-Gruppen und deren jeweilige Gruppennummern.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
------------------------------------------------------------------------
--------
1 +ASM0/TOAST/redo01.log
2 +ASM0/TOAST/redo02.log
3 +ASM0/TOAST/redo03.log
```

2. Geben Sie die Größe der Wiederherstellungsprotokolle ein.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Erstellen Sie für jedes Wiederherstellungsprotokoll eine neue Gruppe, indem Sie die gleiche Größe wie die aktuelle Wiederherstellungsprotokollgruppe verwenden, die den neuen Speicherort des Dateisystems verwendet.

```
RMAN> alter database add logfile '/logs/TOAST/redo/log00.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log01.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log02.rdo' size
52428800;
Statement processed
```

4. Entfernen Sie die alten Logfile-Gruppen, die sich noch im vorherigen Speicher befinden.

```
RMAN> alter database drop logfile group 4;
Statement processed
RMAN> alter database drop logfile group 5;
Statement processed
RMAN> alter database drop logfile group 6;
Statement processed
```

5. Wenn ein Fehler auftritt, der das Löschen eines aktiven Protokolls blockiert, erzwingen Sie einen Switch zum nächsten Protokoll, um die Sperre freizugeben und einen globalen Kontrollpunkt zu erzwingen. Ein

Beispiel ist unten dargestellt. Der Versuch, die Logfile-Gruppe 3, die sich am alten Speicherort befand, zu löschen, wurde abgelehnt, da noch aktive Daten in dieser Logdatei vorhanden waren. Eine Protokollarchivierung, gefolgt von einem Kontrollpunkt, ermöglicht das Löschen von Logdateien.

```
RMAN> alter database drop logfile group 4;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 4 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 4 thread 1:
'+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 4;
Statement processed
```

6. Überprüfen Sie die Umgebung, um sicherzustellen, dass alle standortbasierten Parameter aktualisiert werden.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

7. Das folgende Skript zeigt, wie Sie diesen Prozess vereinfachen können.

```
[root@jfsc1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
TOAST redo logs:
/logs/TOAST/redo/log00.rdo
/logs/TOAST/redo/log01.rdo
/logs/TOAST/redo/log02.rdo
TOAST temp datafiles:
/oradata/TOAST/temp01.dbf
TOAST spfile
spfile                                  string
/orabin/product/12.1.0/dbhome_
                                                      1/dbs/spfileTOAST.ora
TOAST key parameters
control_files /logs/TOAST/arch/control01.ctl,
/logs/TOAST/redo/control02.ctl
log_archive_dest_1 LOCATION=/logs/TOAST/arch
```

8. Wenn die ASM-Datenträgergruppen vollständig evakuiert wurden, können sie jetzt mit abgehängt werden `asmcmd`. In vielen Fällen können Dateien, die zu anderen Datenbanken oder der ASM-Datei spfile/passwd gehören, weiterhin vorhanden sein.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Bereinigung der Datendatei

Der Migrationsprozess kann je nach Verwendung von Oracle RMAN zu Datendateien mit langer oder kryptischer Syntax führen. Im hier gezeigten Beispiel wurde das Backup mit dem Dateiformat von durchgeführt `/oradata/TOAST/%U`. `%U` Gibt an, dass RMAN für jede Datendatei einen eindeutigen Standardnamen erstellen sollte. Das Ergebnis ist ähnlich wie im folgenden Text dargestellt. Die traditionellen Namen der Datendateien sind in die Namen eingebettet. Dies kann mithilfe des in dargestellten skriptgesteuerten Ansatzes bereinigt werden "Bereinigung der ASM-Migration".

```
[root@jfsc1 current]# ./fixuniquenames.pl TOAST
#sqlplus Commands
shutdown immediate;
startup mount;
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/system.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/sysaux.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt /oradata/TOAST/undotbs1.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
/oradata/TOAST/users.dbf
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSTEM_FNO-1_01r5fhjg' to '/oradata/TOAST/system.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSAUX_FNO-2_02r5fhjo' to '/oradata/TOAST/sysaux.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
UNDOTBS1_FNO-3_03r5fhjt' to '/oradata/TOAST/undotbs1.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
USERS_FNO-4_05r5fhk6' to '/oradata/TOAST/users.dbf';
alter database open;
```

Oracle ASM-Ausgleich

Wie bereits erläutert, kann eine Oracle ASM-Festplattengruppe mithilfe des Ausgleichs transparent auf ein neues Storage-System migriert werden. Zusammenfassend ist zu sagen, dass beim Ausbalancieren der vorhandenen LUN-Gruppe LUNs gleicher Größe hinzugefügt werden müssen, gefolgt von einem Drop-Vorgang der vorherigen LUN. Oracle ASM verlagert die zugrunde liegenden Daten automatisch in einem optimalen Layout auf neuen Speicher und gibt dann die alten LUNs nach Abschluss frei.

Der Migrationsprozess nutzt effiziente sequenzielle I/O-Vorgänge und führt im Allgemeinen keine Performance-Unterbrechung durch. Bei Bedarf kann die Migrationsrate jedoch gedrosselt werden.

Identifizieren Sie die zu migrierenden Daten

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NEWDATA_0003 1 10240 /dev/mapper/3600a098038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a098038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a098038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a098038303537762b47594c315862 MEMBER
SQL> select group_number||' '||name from v$asm_diskgroup;
1 NEWDATA
```

Erstellen neuer LUNs

Erstellen Sie neue LUNs gleicher Größe und legen Sie die Mitgliedschaft für Benutzer und Gruppen nach Bedarf fest. Die LUNs sollten als angezeigt werden `CANDIDATE` Festplatten.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
 0 0 /dev/mapper/3600a098038303537762b47594c31586b CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315869 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315858 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c31586a CANDIDATE
NEWDATA_0003 1 10240 /dev/mapper/3600a098038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a098038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a098038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a098038303537762b47594c315862 MEMBER
```

Neue LUNS hinzufügen

Während die Add- und Drop-Vorgänge zusammen ausgeführt werden können, ist es in der Regel einfacher, neue LUNs in zwei Schritten hinzuzufügen. Fügen Sie zunächst die neuen LUNs der Festplattengruppe hinzu. Dieser Schritt führt dazu, dass die Hälfte der Extents von den aktuellen ASM-LUNs auf die neuen LUNs migriert wird.

Die Ausgleichskraft gibt die Rate an, mit der Daten übertragen werden. Je höher die Zahl, desto höher ist die Parallelität der Datenübertragung. Die Migration erfolgt mit effizienten sequenziellen I/O-Vorgängen, die wahrscheinlich keine Performance-Probleme verursachen. Auf Wunsch kann die Ausgleichskraft einer laufenden Migration jedoch mit dem angepasst werden `alter diskgroup [name] rebalance power [level]` Befehl. Für typische Migrationen wird der Wert 5 verwendet.

```
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586b' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315869' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315858' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586a' rebalance power 5;
Diskgroup altered.
```

Überwachen Sie den Betrieb

Ein Ausgleichsoperation kann auf verschiedene Weise überwacht und verwaltet werden. Für dieses Beispiel haben wir den folgenden Befehl verwendet.

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Nach Abschluss der Migration werden keine Vorgänge zur Ausbalancierung gemeldet.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Alte LUNs ablegen

Die Migration ist nun zur Hälfte abgeschlossen. Einige grundlegende Performance-Tests stellen sicher, dass die Umgebung sich in einem ordnungsgemäßen Zustand befindet. Nach Bestätigung können die verbleibenden Daten durch Löschen der alten LUNs verschoben werden. Beachten Sie, dass dies nicht zur sofortigen Freigabe der LUNs führt. Der Drop-Vorgang signalisiert Oracle ASM, die Extents zuerst zu verschieben und dann die LUN freizugeben.

```
sqlplus / as sysasm
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0000 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0001 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0002 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0003 rebalance power 5;
Diskgroup altered.
```

Überwachen Sie den Betrieb

Der Ausgleichsoperation kann auf verschiedene Weise überwacht und verwaltet werden. Für dieses Beispiel haben wir den folgenden Befehl verwendet:

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Nach Abschluss der Migration werden keine Vorgänge zur Ausbalancierung gemeldet.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Entfernen Sie alte LUNs

Bevor Sie die alten LUNs aus der Laufwerksgruppe entfernen, sollten Sie den Header-Status einer letzten Prüfung entnehmen. Nachdem eine LUN aus ASM freigegeben wurde, wird kein Name mehr aufgeführt, und der Kopfzeilenstatus wird als aufgeführt `FORMER`. Dies bedeutet, dass diese LUNs sicher aus dem System entfernt werden können.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NAME||''||GROUP_NUMBER||''||TOTAL_MB||''||PATH||''||HEADER_STATUS
------------------------------------------------------------------------
------
 0 0 /dev/mapper/3600a098038303537762b47594c315863 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315864 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315861 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315862 FORMER
NEWDATA_0005 1 10240 /dev/mapper/3600a098038303537762b47594c315869 MEMBER
NEWDATA_0007 1 10240 /dev/mapper/3600a098038303537762b47594c31586a MEMBER
NEWDATA_0004 1 10240 /dev/mapper/3600a098038303537762b47594c31586b MEMBER
NEWDATA_0006 1 10240 /dev/mapper/3600a098038303537762b47594c315858 MEMBER
8 rows selected.
```

LVM-Migration

Das hier vorgestellte Verfahren zeigt die Prinzipien einer LVM-basierten Migration einer Volume-Gruppe namens `datavg`. Die Beispiele stammen aus Linux LVM, die Prinzipien gelten jedoch gleichermaßen für AIX, HP-UX und VxVM. Die genauen Befehle können variieren.

1. Identifizieren Sie die LUNs, die sich derzeit im befinden `datavg` Volume-Gruppe.

```
[root@host1 ~]# pvdisplay -C | grep datavg
  /dev/mapper/3600a098038303537762b47594c31582f datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31585a datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c315859 datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31586c datavg lvm2 a--  10.00g
10.00g
```

2. Erstellen Sie neue LUNs mit derselben oder einer etwas größeren physischen Größe und definieren Sie sie als physische Volumes.

```
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315864
  Physical volume "/dev/mapper/3600a098038303537762b47594c315864"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315863
  Physical volume "/dev/mapper/3600a098038303537762b47594c315863"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315862
  Physical volume "/dev/mapper/3600a098038303537762b47594c315862"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315861
  Physical volume "/dev/mapper/3600a098038303537762b47594c315861"
successfully created
```

3. Fügen Sie die neuen Volumes zur Volume-Gruppe hinzu.

```
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315864
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315863
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315862
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315861
  Volume group "datavg" successfully extended
```

4. Stellen Sie das aus `pvmove` Befehl, um die Extents jeder aktuellen LUN in die neue LUN zu verschieben. Der `-i [seconds]` Argument überwacht den Fortschritt des Vorgangs.

```
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31582f
/dev/mapper/3600a098038303537762b47594c315864
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 14.2%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 28.4%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 42.5%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 57.1%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 72.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 87.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31585a
/dev/mapper/3600a098038303537762b47594c315863
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 14.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 29.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 44.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 60.1%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 75.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 90.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c315859
/dev/mapper/3600a098038303537762b47594c315862
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 14.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 29.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 45.5%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 61.1%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 76.6%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 91.7%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31586c
/dev/mapper/3600a098038303537762b47594c315861
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 15.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 30.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 46.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 61.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 77.2%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 92.3%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 100.0%
```

5. Wenn dieser Vorgang abgeschlossen ist, löschen Sie die alten LUNs aus der Volume-Gruppe mithilfe von `vgreduce` Befehl. Wenn die LUN erfolgreich war, kann sie jetzt sicher aus dem System entfernt werden.

```
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31582f
Removed "/dev/mapper/3600a098038303537762b47594c31582f" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31585a
  Removed "/dev/mapper/3600a098038303537762b47594c31585a" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c315859
  Removed "/dev/mapper/3600a098038303537762b47594c315859" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31586c
  Removed "/dev/mapper/3600a098038303537762b47594c31586c" from volume
group "datavg"
```

Import fremder LUNs

Planung

Die Verfahren zur Migration von SAN-Ressourcen mit FLI sind in NetApp dokumentiert "ONTAP Dokumentation zum Importieren fremder LUNs".

Aus Sicht der Datenbank und des Hosts sind keine besonderen Schritte erforderlich. Nachdem die FC-Zonen aktualisiert wurden und die LUNs auf ONTAP verfügbar werden, sollte die LVM in der Lage sein, die LVM-Metadaten von den LUNs zu lesen. Außerdem sind die Volume-Gruppen ohne weitere Konfigurationsschritte einsatzbereit. In seltenen Fällen können Umgebungen Konfigurationsdateien enthalten, die hartcodiert waren und Verweise auf das vorherige Storage-Array enthalten. Zum Beispiel ein Linux-System, das enthalten `/etc/multipath.conf` Regeln, die auf einen WWN eines bestimmten Geräts verwiesen haben, müssen aktualisiert werden, um die von FLI eingeführten Änderungen wiederzugeben.

> Informationen zu unterstützten Konfigurationen finden Sie in der NetApp Kompatibilitätsmatrix. Falls Ihr System nicht im Lieferumfang enthalten ist, wenden Sie sich an Ihren NetApp Ansprechpartner.

Dieses Beispiel zeigt die Migration von ASM- und LVM-LUNs, die auf einem Linux-Server gehostet werden. FLI wird auf anderen Betriebssystemen unterstützt, und obwohl die Host-seitigen Befehle unterschiedlich sein können, sind die Prinzipien identisch, und die ONTAP-Verfahren sind identisch.

LVM-LUNs identifizieren

Der erste Schritt zur Vorbereitung besteht darin, die zu migrierenden LUNs zu identifizieren. In dem hier gezeigten Beispiel werden zwei SAN-basierte Dateisysteme in gemountet `/orabin` Und `/backups`.

```
[root@host1 ~]# df -k
Filesystem                          1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root                52403200   8811464  43591736  17% /
devtmpfs                             65882776         0  65882776   0% /dev
...
fas8060-nfs-public:/install         199229440 119368128  79861312  60%
/install
/dev/mapper/sanvg-lvorabin           20961280  12348476   8612804  59%
/orabin
/dev/mapper/sanvg-lvbackups          73364480  62947536  10416944  86%
/backups
```

Der Name der Volume-Gruppe kann aus dem Gerätenamen extrahiert werden, der das Format (Name der Volume-Gruppe)-(Name des logischen Volumes) verwendet. In diesem Fall wird die Volume-Gruppe aufgerufen `sanvg`.

Der `pvdisplay` Mit dem Befehl können Sie die LUNs identifizieren, die diese Volume-Gruppe unterstützen. In diesem Fall sind 10 LUNs vorhanden `sanvg` Volume-Gruppe.

```
[root@host1 ~]# pvdisplay -C -o pv_name,pv_size,pv_fmt,vg_name
  PV                                                PSize   VG
  /dev/mapper/3600a0980383030445424487556574266  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574267  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574268  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574269  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426a  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426b  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426c  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426d  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426e  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426f  10.00g sanvg
  /dev/sda2                                       278.38g rhel
```

ASM-LUNs identifizieren

ASM-LUNs müssen ebenfalls migriert werden. Um die Anzahl der LUNs und LUN-Pfade von sqlplus als sysasm-Benutzer zu erhalten, führen Sie den folgenden Befehl aus:

```
SQL> select path||' '||os_mb from v$asm_disk;
PATH||''||OS_MB
--------------------------------------------------------------------------
------
/dev/oracleasm/disks/ASM0 10240
/dev/oracleasm/disks/ASM9 10240
/dev/oracleasm/disks/ASM8 10240
/dev/oracleasm/disks/ASM7 10240
/dev/oracleasm/disks/ASM6 10240
/dev/oracleasm/disks/ASM5 10240
/dev/oracleasm/disks/ASM4 10240
/dev/oracleasm/disks/ASM1 10240
/dev/oracleasm/disks/ASM3 10240
/dev/oracleasm/disks/ASM2 10240
10 rows selected.
SQL>
```

Änderungen am FC-Netzwerk

Die aktuelle Umgebung enthält 20 zu migrierende LUNs. Aktualisieren Sie das aktuelle SAN, damit ONTAP auf die aktuellen LUNs zugreifen kann. Daten werden noch nicht migriert, aber ONTAP muss die Konfigurationsinformationen der aktuellen LUNs lesen, um das neue Zuhause für diese Daten zu erstellen.

Mindestens ein HBA-Port auf dem All Flash FAS/FAS System muss als Initiator-Port konfiguriert sein. Zudem müssen die FC-Zonen aktualisiert werden, damit ONTAP auf die LUNs auf dem fremden Storage Array zugreifen können. Bei einigen Speicher-Arrays ist die LUN-Maskierung konfiguriert, wodurch WWNs auf eine bestimmte LUN zugreifen können. In diesen Fällen muss die LUN-Maskierung ebenfalls aktualisiert werden, um Zugriff auf die ONTAP-WWNs zu gewähren.

Nach Abschluss dieses Schritts sollte ONTAP in der Lage sein, das fremde Speicher-Array mit dem anzuzeigen `storage array show` Befehl. Das Schlüsselfeld, das zurückgegeben wird, ist das Präfix, das zur Identifizierung der fremden LUN auf dem System verwendet wird. Im folgenden Beispiel werden die LUNs auf dem Fremdarray angezeigt `FOREIGN_1` Wird in ONTAP mit dem Präfix von angezeigt `FOR-1`.

Identifizierung von Fremdarrays

```
Cluster01::> storage array show -fields name,prefix
name          prefix
------------- ------
FOREIGN_1     FOR-1
Cluster01::>
```

Identifizierung fremder LUNs

Die LUNs können durch Bestehen des aufgelistet werden `array-name` Bis zum `storage disk show` Befehl. Die zurückgegebenen Daten werden während des Migrationsvorgangs mehrfach referenziert.

```
Cluster01::> storage disk show -array-name FOREIGN_1 -fields disk,serial
disk     serial-number
-------- -------------
FOR-1.1  800DT$HuVWBX
FOR-1.2  800DT$HuVWBZ
FOR-1.3  800DT$HuVWBW
FOR-1.4  800DT$HuVWBY
FOR-1.5  800DT$HuVWB/
FOR-1.6  800DT$HuVWBa
FOR-1.7  800DT$HuVWBd
FOR-1.8  800DT$HuVWBb
FOR-1.9  800DT$HuVWBc
FOR-1.10 800DT$HuVWBe
FOR-1.11 800DT$HuVWBf
FOR-1.12 800DT$HuVWBg
FOR-1.13 800DT$HuVWBi
FOR-1.14 800DT$HuVWBh
FOR-1.15 800DT$HuVWBj
FOR-1.16 800DT$HuVWBk
FOR-1.17 800DT$HuVWBm
FOR-1.18 800DT$HuVWBl
FOR-1.19 800DT$HuVWBo
FOR-1.20 800DT$HuVWBn
20 entries were displayed.
Cluster01::>
```

Registrieren Sie LUNs für Fremdarrays als Importkandidaten

Die ausländischen LUNs werden zunächst als jeder bestimmte LUN-Typ klassifiziert. Bevor Daten importiert werden können, müssen die LUNs als fremd gekennzeichnet werden und daher als Kandidat für den Importprozess. Um diesen Schritt abzuschließen, geben Sie die Seriennummer an den weiter `storage disk modify` Wie im folgenden Beispiel gezeigt. Beachten Sie, dass bei diesem Prozess nur die LUN als fremd innerhalb von ONTAP markiert wird. Es werden keine Daten auf die fremde LUN selbst geschrieben.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign true
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign true
Cluster01::*>
```

Erstellung von Volumes zum Hosten migrierter LUNs

Ein Volume ist erforderlich, um die migrierten LUNs zu hosten. Die genaue Volume-Konfiguration hängt von der Planung der Nutzung von ONTAP Funktionen ab. In diesem Beispiel werden die ASM-LUNs in einem Volume platziert und die LVM-LUNs in einem zweiten Volume platziert. Auf diese Weise können Sie die LUNs als unabhängige Gruppen managen, beispielsweise für Tiering, die Erstellung von Snapshots oder die Einstellung von QoS-Kontrollen.

Stellen Sie die ein `snapshot-policy `to `none`. Der Migrationsprozess kann sehr viel Datenfluktuation beinhalten. Daher kann es zu einem starken Anstieg des Platzverbrauchs kommen, wenn Snapshots versehentlich erstellt werden, weil unerwünschte Daten in den Snapshots erfasst werden.

```
Cluster01::> volume create -volume new_asm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1152] Job succeeded: Successful
Cluster01::> volume create -volume new_lvm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1153] Job succeeded: Successful
Cluster01::>
```

Erstellen Sie ONTAP-LUNs

Nach der Erstellung der Volumes müssen die neuen LUNs erstellt werden. Normalerweise erfordert die Erstellung einer LUN, dass der Benutzer Informationen wie die LUN-Größe angeben muss. In diesem Fall wird jedoch das Argument für eine fremde Festplatte an den Befehl übergeben. Infolgedessen repliziert ONTAP die aktuellen LUN-Konfigurationsdaten von der angegebenen Seriennummer. Außerdem werden die LUN-Geometrie und Partitionstabellen-Daten verwendet, um die LUN-Ausrichtung anzupassen und eine optimale Performance herzustellen.

In diesem Schritt müssen die Seriennummern mit dem Fremdarray verglichen werden, um sicherzustellen, dass die richtige fremde LUN mit der richtigen neuen LUN abgeglichen wird.

```
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN0 -ostype
linux -foreign-disk 800DT$HuVWBW
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN1 -ostype
linux -foreign-disk 800DT$HuVWBX
Created a LUN of size 10g (10737418240)
...
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN8 -ostype
linux -foreign-disk 800DT$HuVWBn
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN9 -ostype
linux -foreign-disk 800DT$HuVWBo
Created a LUN of size 10g (10737418240)
```

Erstellen Sie Importbeziehungen

Die LUNs wurden jetzt erstellt, sind aber nicht als Replikationsziel konfiguriert. Bevor dieser Schritt durchgeführt werden kann, müssen die LUNs zunächst in den Offline-Modus versetzt werden. Dieser zusätzliche Schritt dient dem Schutz von Daten vor Benutzerfehlern. Wenn ONTAP die Durchführung einer Migration auf einer Online-LUN zulässt, besteht das Risiko, dass durch einen typografischen Fehler aktive Daten überschrieben werden. Durch den zusätzlichen Schritt, den Benutzer zum ersten Mal offline zu schalten, wird überprüft, ob die richtige Ziel-LUN als Migrationsziel verwendet wird.

```
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN0
Warning: This command will take LUN "/vol/new_asm/LUN0" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN1
Warning: This command will take LUN "/vol/new_asm/LUN1" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
...
Warning: This command will take LUN "/vol/new_lvm/LUN8" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_lvm/LUN9
Warning: This command will take LUN "/vol/new_lvm/LUN9" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
```

Nachdem die LUNs offline sind, können Sie die Importbeziehung wiederherstellen, indem Sie die Seriennummer der fremden LUN an den übergeben `lun import create` Befehl.

```
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN0
-foreign-disk 800DT$HuVWBW
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN1
-foreign-disk 800DT$HuVWBX
...
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN8
-foreign-disk 800DT$HuVWBn
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN9
-foreign-disk 800DT$HuVWBo
Cluster01::*>
```

Nachdem alle Importbeziehungen eingerichtet sind, können die LUNs wieder online geschaltet werden.

```
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN9
Cluster01::*>
```

Erstellen einer Initiatorgruppe

Eine Initiatorgruppe (Initiatorgruppe) ist Teil der ONTAP LUN-Masking-Architektur. Auf eine neu erstellte LUN kann nur dann zugegriffen werden, wenn einem Host der erste Zugriff gewährt wurde. Dazu wird eine Initiatorgruppe erstellt, die entweder die FC-WWNs oder iSCSI-Initiatornamen auflistet, denen Zugriff gewährt werden soll. Zum Zeitpunkt der Erstellung dieses Berichts wurde FLI nur für FC LUNs unterstützt. Die Konvertierung in iSCSI nach der Migration ist jedoch eine einfache Aufgabe, wie in dargestellt "Protokollkonvertierung".

In diesem Beispiel wird eine Initiatorgruppe erstellt, die zwei WWNs enthält, die den beiden auf dem HBA des Hosts verfügbaren Ports entsprechen.

```
Cluster01::*> igroup create linuxhost -protocol fcp -ostype linux
-initiator 21:00:00:0e:1e:16:63:50 21:00:00:0e:1e:16:63:51
```

Ordnen Sie neue LUNs dem Host zu

Nach der Erstellung der Initiatorgruppe werden die LUNs dann der definierten Initiatorgruppe zugeordnet. Diese LUNs sind nur für die WWNs dieser Initiatorgruppe verfügbar. NetApp geht in dieser Phase des Migrationsprozesses davon aus, dass der Host nicht auf ONTAP abgegrenzt wurde. Dies ist wichtig, denn wenn der Host gleichzeitig auf das fremde Array und das neue ONTAP-System begrenzt ist, besteht das Risiko, dass LUNs mit derselben Seriennummer auf jedem Array erkannt werden können. Diese Situation kann zu Fehlfunktionen des Multipfad-Funktionszubers oder zu Schäden an Daten führen.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
Cluster01::*>
```

Umstellung

Aufgrund der Notwendigkeit, die FC-Netzwerkkonfiguration zu ändern, sind Unterbrechungen beim Import fremder LUNs unvermeidbar. Die Unterbrechung muss

jedoch nicht viel länger dauern als die Zeit, die für den Neustart der Datenbankumgebung und die Aktualisierung des FC-Zoning für die Umstellung der Host-FC-Konnektivität von der fremden LUN auf ONTAP erforderlich ist.

Dieser Prozess lässt sich wie folgt zusammenfassen:

1. Legen Sie alle LUN-Aktivitäten auf den fremden LUNs still.
2. Umleiten von Host-FC-Verbindungen zum neuen ONTAP-System
3. Starten Sie den Importvorgang.
4. Ermitteln Sie die LUNs neu.
5. Starten Sie die Datenbank neu.

Sie müssen nicht warten, bis der Migrationsprozess abgeschlossen ist. Sobald die Migration einer bestimmten LUN beginnt, ist sie auf ONTAP verfügbar und kann Daten bereitstellen, während der Datenkopievorgang fortgesetzt wird. Alle Lesevorgänge werden an die fremde LUN weitergeleitet, und alle Schreibvorgänge werden synchron auf beide Arrays geschrieben. Der Kopiervorgang läuft sehr schnell ab und der Overhead bei der Umleitung des FC-Datenverkehrs ist minimal. Die Auswirkungen auf die Performance sollten daher kurzlebig und minimal sein. Wenn Bedenken bestehen, können Sie den Neustart der Umgebung verzögern, bis der Migrationsprozess abgeschlossen ist und die Importbeziehungen gelöscht wurden.

Datenbank herunterfahren

Der erste Schritt bei der Stilllegung der Umgebung in diesem Beispiel ist das Herunterfahren der Datenbank.

```
[oracle@host1 bin]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base remains unchanged with value /orabin
[oracle@host1 bin]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, Automatic Storage Management, OLAP, Advanced
Analytics
and Real Application Testing options
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
SQL>
```

Netzdienste herunterfahren

Zu den migrierten SAN-basierten Dateisystemen gehören auch die Oracle ASM-Services. Um die zugrunde liegenden LUNs stilllegen zu können, müssen die Dateisysteme getrennt werden. Dies bedeutet wiederum, dass alle Prozesse mit offenen Dateien auf diesem Dateisystem angehalten werden.

```
[oracle@host1 bin]$ ./crsctl stop has -f
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'host1'
CRS-2673: Attempting to stop 'ora.evmd' on 'host1'
CRS-2673: Attempting to stop 'ora.DATA.dg' on 'host1'
CRS-2673: Attempting to stop 'ora.LISTENER.lsnr' on 'host1'
CRS-2677: Stop of 'ora.DATA.dg' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.asm' on 'host1'
CRS-2677: Stop of 'ora.LISTENER.lsnr' on 'host1' succeeded
CRS-2677: Stop of 'ora.evmd' on 'host1' succeeded
CRS-2677: Stop of 'ora.asm' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.cssd' on 'host1'
CRS-2677: Stop of 'ora.cssd' on 'host1' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'host1' has completed
CRS-4133: Oracle High Availability Services has been stopped.
[oracle@host1 bin]$
```

Entfernen Sie Dateisysteme

Wenn alle Prozesse heruntergefahren werden, ist der umount-Vorgang erfolgreich. Wenn die Berechtigung verweigert wird, muss es einen Prozess mit einer Sperre auf dem Dateisystem geben. Der `fuser` Der Befehl kann bei der Identifizierung dieser Prozesse helfen.

```
[root@host1 ~]# umount /orabin
[root@host1 ~]# umount /backups
```

Deaktivieren Sie Volume-Gruppen

Nachdem alle Dateisysteme in einer bestimmten Volume-Gruppe getrennt wurden, kann die Volume-Gruppe deaktiviert werden.

```
[root@host1 ~]# vgchange --activate n sanvg
  0 logical volume(s) in volume group "sanvg" now active
[root@host1 ~]#
```

Änderungen am FC-Netzwerk

Die FC-Zonen können jetzt aktualisiert werden, um den gesamten Zugriff vom Host auf das fremde Array zu entfernen und den Zugriff auf ONTAP zu ermöglichen.

Importvorgang starten

Um die LUN-Importprozesse zu starten, führen Sie den aus `lun import start` Befehl.

```
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN0
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN1
...
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN8
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN9
Cluster01::lun import*>
```

Überwachen Sie den Importfortschritt

Der Importvorgang kann mit dem überwacht werden `lun import show` Befehl. Wie unten dargestellt, läuft der Import aller 20 LUNs, was bedeutet, dass die Daten jetzt über ONTAP zugänglich sind, obwohl der Kopiervorgang noch fortschreitet.

```
Cluster01::lun import*> lun import show -fields path,percent-complete
vserver   foreign-disk path              percent-complete
--------- ------------ ----------------- ----------------
vserver1  800DT$HuVWB/ /vol/new_asm/LUN4 5
vserver1  800DT$HuVWBW /vol/new_asm/LUN0 5
vserver1  800DT$HuVWBX /vol/new_asm/LUN1 6
vserver1  800DT$HuVWBY /vol/new_asm/LUN2 6
vserver1  800DT$HuVWBZ /vol/new_asm/LUN3 5
vserver1  800DT$HuVWBa /vol/new_asm/LUN5 4
vserver1  800DT$HuVWBb /vol/new_asm/LUN6 4
vserver1  800DT$HuVWBc /vol/new_asm/LUN7 4
vserver1  800DT$HuVWBd /vol/new_asm/LUN8 4
vserver1  800DT$HuVWBe /vol/new_asm/LUN9 4
vserver1  800DT$HuVWBf /vol/new_lvm/LUN0 5
vserver1  800DT$HuVWBg /vol/new_lvm/LUN1 4
vserver1  800DT$HuVWBh /vol/new_lvm/LUN2 4
vserver1  800DT$HuVWBi /vol/new_lvm/LUN3 3
vserver1  800DT$HuVWBj /vol/new_lvm/LUN4 3
vserver1  800DT$HuVWBk /vol/new_lvm/LUN5 3
vserver1  800DT$HuVWBl /vol/new_lvm/LUN6 4
vserver1  800DT$HuVWBm /vol/new_lvm/LUN7 3
vserver1  800DT$HuVWBn /vol/new_lvm/LUN8 2
vserver1  800DT$HuVWBo /vol/new_lvm/LUN9 2
20 entries were displayed.
```

Wenn Sie einen Offline-Prozess benötigen, verzögern Sie die Neuermittlung oder den Neustart von Diensten, bis der `lun import show` Befehl anzeigt, dass alle Migration erfolgreich und abgeschlossen ist. Anschließend können Sie den Migrationsprozess wie unter beschrieben abschließen"Import fremder LUNs –

Abschluss".

Wenn Sie eine Online-Migration benötigen, fahren Sie mit der Neuerkennung der LUNs in ihrem neuen Zuhause fort, und führen Sie die Dienste aus.

Nach SCSI-Geräteänderungen suchen

In den meisten Fällen besteht die einfachste Möglichkeit, neue LUNs neu zu ermitteln, darin, den Host neu zu starten. Dadurch werden alte veraltete Geräte automatisch entfernt, alle neuen LUNs ordnungsgemäß erkannt und verbundene Geräte wie Multipathing-Geräte erstellt. Das Beispiel zeigt einen vollständig online-Prozess zu Demonstrationszwecken.

Achtung: Bevor Sie einen Host neu starten, stellen Sie sicher, dass alle Einträge in sind `/etc/fstab` Diese Referenz migrierte SAN-Ressourcen werden kommentiert. Wenn dies nicht durchgeführt wird und Probleme mit dem LUN-Zugriff auftreten, wird das OS möglicherweise nicht gebootet. Diese Situation beschädigt Daten nicht. Es kann jedoch sehr unbequem sein, in den Rettungsmodus oder einen ähnlichen Modus zu starten und die zu korrigieren `/etc/fstab` Damit das OS gebootet werden kann, um die Fehlerbehebung zu ermöglichen.

Die LUNs auf der in diesem Beispiel verwendeten Linux-Version können erneut mit dem gescannt werden `rescan-scsi-bus.sh` Befehl. Wenn der Befehl erfolgreich war, sollte jeder LUN-Pfad in der Ausgabe angezeigt werden. Die Ausgabe kann schwer zu interpretieren sein, wenn die Zoning- und igroup-Konfiguration korrekt war, sollten viele LUNs scheinen, die eine enthalten `NETAPP` Anbieterzeichenfolge.

```
[root@host1 /]# rescan-scsi-bus.sh
Scanning SCSI subsystem for new devices
Scanning host 0 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 0 2 0 0 ...
OLD: Host: scsi0 Channel: 02 Id: 00 Lun: 00
      Vendor: LSI      Model: RAID SAS 6G 0/1  Rev: 2.13
      Type:   Direct-Access                    ANSI SCSI revision: 05
Scanning host 1 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 1 0 0 0 ...
OLD: Host: scsi1 Channel: 00 Id: 00 Lun: 00
      Vendor: Optiarc  Model: DVD RW AD-7760H  Rev: 1.41
      Type:   CD-ROM                           ANSI SCSI revision: 05
Scanning host 2 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 3 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 4 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 5 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 6 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 7 for  all SCSI target IDs, all LUNs
 Scanning for device 7 0 0 10 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 10
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 11 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 11
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 12 ...
...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 18
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 9 0 1 19 ...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 19
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
0 new or changed device(s) found.
0 remapped or resized device(s) found.
0 device(s) removed.
```

Überprüfen Sie auf Multipath-Geräte

Der LUN-Erkennungsprozess löst auch die Wiederherstellung von Multipath-Geräten aus, der Linux-Multipathing-Treiber hat jedoch bekanntermaßen gelegentlich Probleme. Die Ausgabe von `multipath - ll` Sollte überprüft werden, um sicherzustellen, dass die Ausgabe wie erwartet aussieht. Die folgende Ausgabe zeigt beispielsweise Multipath-Geräte, die mit einem verknüpft sind `NETAPP` Anbieterzeichenfolge. Jedes Gerät verfügt über vier Pfade, wobei zwei mit einer Priorität von 50 und zwei mit einer Priorität von 10. Obwohl die

genaue Ausgabe mit verschiedenen Versionen von Linux variieren kann, sieht diese Ausgabe wie erwartet aus.

Überprüfen Sie anhand der Dokumentation der Host-Dienstprogramme die Version von Linux, die Sie verwenden `/etc/multipath.conf` Die Einstellungen sind korrekt.

```
[root@host1 /]# multipath -ll
3600a098038303558735d493762504b36 dm-5 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:4  sdat 66:208 active ready running
| `- 9:0:1:4  sdbn 68:16  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:4  sdf  8:80   active ready running
  `- 9:0:0:4  sdz  65:144 active ready running
3600a098038303558735d493762504b2d dm-10 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:8  sdax 67:16  active ready running
| `- 9:0:1:8  sdbr 68:80  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:8  sdj  8:144  active ready running
  `- 9:0:0:8  sdad 65:208 active ready running
...
3600a098038303558735d493762504b37 dm-8 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:5  sdau 66:224 active ready running
| `- 9:0:1:5  sdbo 68:32  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:5  sdg  8:96   active ready running
  `- 9:0:0:5  sdaa 65:160 active ready running
3600a098038303558735d493762504b4b dm-22 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:19 sdbi 67:192 active ready running
| `- 9:0:1:19 sdcc 69:0   active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:19 sdu  65:64  active ready running
  `- 9:0:0:19 sdao 66:128 active ready running
```

Reaktivieren Sie die LVM-Volume-Gruppe

Wenn die LVM-LUNs ordnungsgemäß erkannt wurden, wird das angezeigt `vgchange --activate y` Befehl sollte erfolgreich sein. Dies ist ein gutes Beispiel für den Nutzen eines logischen Volume-Managers. Eine Änderung des WWN einer LUN oder auch einer Seriennummer ist unwichtig, da die Metadaten der Volume-Gruppe auf die LUN selbst geschrieben werden.

Das Betriebssystem hat die LUNs gescannt und eine kleine Menge an auf die LUN geschriebenen Daten ermittelt, die sie als physisches Volume des identifizieren `sanvg volumegroup`. Anschließend wurden alle erforderlichen Geräte erstellt. Sie müssen nur die Volume-Gruppe erneut aktivieren.

```
[root@host1 /]# vgchange --activate y sanvg
  Found duplicate PV fpCzdLTuKfy2xDZjai1NliJh3TjLUBiT: using
/dev/mapper/3600a098038303558735d493762504b46 not /dev/sdp
  Using duplicate PV /dev/mapper/3600a098038303558735d493762504b46 from
subsystem DM, ignoring /dev/sdp
  2 logical volume(s) in volume group "sanvg" now active
```

Dateisysteme neu einbinden

Nachdem die Volume-Gruppe wieder aktiviert wurde, können die Dateisysteme mit allen ursprünglichen Daten gemountet werden. Wie bereits erwähnt, sind die Dateisysteme voll funktionsfähig, selbst wenn die Datenreplikation in der Back-Gruppe weiterhin aktiv ist.

```
[root@host1 /]# mount /orabin
[root@host1 /]# mount /backups
[root@host1 /]# df -k
Filesystem                          1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root                52403200   8837100  43566100  17% /
devtmpfs                             65882776         0  65882776   0% /dev
tmpfs                                 6291456        84   6291372   1%
/dev/shm
tmpfs                                65898668      9884  65888784   1% /run
tmpfs                                65898668         0  65898668   0%
/sys/fs/cgroup
/dev/sda1                              505580    224828    280752  45% /boot
fas8060-nfs-public:/install          199229440 119368256  79861184  60%
/install
fas8040-nfs-routable:/snapomatic      9961472     30528   9930944   1%
/snapomatic
tmpfs                                13179736        16  13179720   1%
/run/user/42
tmpfs                                13179736         0  13179736   0%
/run/user/0
/dev/mapper/sanvg-lvorabin           20961280  12357456   8603824  59%
/orabin
/dev/mapper/sanvg-lvbackups          73364480  62947536  10416944  86%
/backups
```

Neuscannen für ASM-Geräte

Die ASMlib-Geräte sollten beim erneuten Scannen der SCSI-Geräte neu erkannt worden sein. Die Wiedererkennung kann online überprüft werden, indem ASMlib neu gestartet und anschließend die Datenträger gescannt werden.

Dieser Schritt ist nur für ASM-Konfigurationen relevant, in denen ASMlib verwendet wird.

Achtung: Wenn ASMlib nicht verwendet wird, ist die `/dev/mapper` Geräte sollten automatisch neu erstellt worden sein. Die Berechtigungen sind jedoch möglicherweise nicht korrekt. Sie müssen spezielle Berechtigungen für die zugrunde liegenden Geräte für ASM festlegen, wenn ASMlib nicht vorhanden ist. Dies wird in der Regel durch spezielle Einträge in entweder der erreicht `/etc/multipath.conf` Oder `udev` Regeln oder möglicherweise in beiden Regelsätzen. Diese Dateien müssen möglicherweise aktualisiert werden, um Änderungen in der Umgebung in Bezug auf WWNs oder Seriennummern widerzuspiegeln, um sicherzustellen, dass die ASM-Geräte weiterhin über die richtigen Berechtigungen verfügen.

In diesem Beispiel werden beim Neustart von ASMlib und beim Scannen nach Festplatten die gleichen 10 ASM-LUNs wie in der ursprünglichen Umgebung angezeigt.

```
[root@host1 /]# oracleasm exit
Unmounting ASMlib driver filesystem: /dev/oracleasm
Unloading module "oracleasm": oracleasm
[root@host1 /]# oracleasm init
Loading module "oracleasm": oracleasm
Configuring "oracleasm" to use device physical block size
Mounting ASMlib driver filesystem: /dev/oracleasm
[root@host1 /]# oracleasm scandisks
Reloading disk partitions: done
Cleaning any stale ASM disks...
Scanning system for ASM disks...
Instantiating disk "ASM0"
Instantiating disk "ASM1"
Instantiating disk "ASM2"
Instantiating disk "ASM3"
Instantiating disk "ASM4"
Instantiating disk "ASM5"
Instantiating disk "ASM6"
Instantiating disk "ASM7"
Instantiating disk "ASM8"
Instantiating disk "ASM9"
```

Starten Sie die Grid-Services neu

Da die LVM- und ASM-Geräte jetzt online und verfügbar sind, können die Grid-Dienste neu gestartet werden.

```
[root@host1 /]# cd /orabin/product/12.1.0/grid/bin
[root@host1 bin]# ./crsctl start has
```

Datenbank neu starten

Nach dem Neustart der Netzdienste kann die Datenbank gestartet werden. Möglicherweise müssen Sie einige Minuten warten, bis die ASM-Dienste vollständig verfügbar sind, bevor Sie versuchen, die Datenbank zu starten.

```
[root@host1 bin]# su - oracle
[oracle@host1 ~]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base has been set to /orabin
[oracle@host1 ~]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> startup
ORACLE instance started.
Total System Global Area 3221225472 bytes
Fixed Size                  4502416 bytes
Variable Size            1207962736 bytes
Database Buffers         1996488704 bytes
Redo Buffers               12271616 bytes
Database mounted.
Database opened.
SQL>
```

Abschluss

Aus Host-Sicht ist die Migration abgeschlossen, aber I/O wird weiterhin vom fremden Array bedient, bis die Importbeziehungen gelöscht werden.

Bevor Sie die Beziehungen löschen, müssen Sie bestätigen, dass der Migrationsprozess für alle LUNs abgeschlossen ist.

```
Cluster01::*> lun import show -vserver vserver1 -fields foreign-
disk,path,operational-state
vserver   foreign-disk path              operational-state
--------- ------------ ----------------- -----------------
vserver1 800DT$HuVWB/ /vol/new_asm/LUN4 completed
vserver1 800DT$HuVWBW /vol/new_asm/LUN0 completed
vserver1 800DT$HuVWBX /vol/new_asm/LUN1 completed
vserver1 800DT$HuVWBY /vol/new_asm/LUN2 completed
vserver1 800DT$HuVWBZ /vol/new_asm/LUN3 completed
vserver1 800DT$HuVWBa /vol/new_asm/LUN5 completed
vserver1 800DT$HuVWBb /vol/new_asm/LUN6 completed
vserver1 800DT$HuVWBc /vol/new_asm/LUN7 completed
vserver1 800DT$HuVWBd /vol/new_asm/LUN8 completed
vserver1 800DT$HuVWBe /vol/new_asm/LUN9 completed
vserver1 800DT$HuVWBf /vol/new_lvm/LUN0 completed
vserver1 800DT$HuVWBg /vol/new_lvm/LUN1 completed
vserver1 800DT$HuVWBh /vol/new_lvm/LUN2 completed
vserver1 800DT$HuVWBi /vol/new_lvm/LUN3 completed
vserver1 800DT$HuVWBj /vol/new_lvm/LUN4 completed
vserver1 800DT$HuVWBk /vol/new_lvm/LUN5 completed
vserver1 800DT$HuVWBl /vol/new_lvm/LUN6 completed
vserver1 800DT$HuVWBm /vol/new_lvm/LUN7 completed
vserver1 800DT$HuVWBn /vol/new_lvm/LUN8 completed
vserver1 800DT$HuVWBo /vol/new_lvm/LUN9 completed
20 entries were displayed.
```

Importbeziehungen löschen

Löschen Sie nach Abschluss des Migrationsprozesses die Migrationsbeziehung. Anschließend wird die I/O ausschließlich von den Laufwerken auf ONTAP bedient.

```
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN9
```

Registrierung ausländischer LUNs aufheben

Ändern Sie schließlich die Festplatte, um die zu entfernen `is-foreign` Bezeichnung.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign false
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign false
Cluster01::*>
```

Protokollkonvertierung

Das Ändern des Protokolls für den Zugriff auf eine LUN ist eine gängige Anforderung.

In einigen Fällen ist die Migration der Daten in die Cloud Teil einer Gesamtstrategie. TCP/IP ist das Protokoll der Cloud, und der Wechsel von FC zu iSCSI ermöglicht eine einfachere Migration in verschiedene Cloud-Umgebungen. In anderen Fällen kann iSCSI wünschenswert sein, die gesunkenen Kosten eines IP SAN zu nutzen. Gelegentlich kann eine Migration ein anderes Protokoll als temporäre Maßnahme verwenden. Wenn beispielsweise ein fremdes Array und ONTAP-basierte LUNs nicht auf denselben HBAs koexistieren können, können Sie iSCSI-LUNs verwenden, die lang genug sind, um Daten vom alten Array zu kopieren. Nachdem die alten LUNs aus dem System entfernt wurden, können Sie sie wieder zu FC konvertieren.

Das folgende Verfahren zeigt die Konvertierung von FC zu iSCSI, jedoch gelten die allgemeinen Prinzipien für eine umgekehrte iSCSI- zu FC-Konvertierung.

Installieren Sie den iSCSI-Initiator

Die meisten Betriebssysteme enthalten standardmäßig einen Software-iSCSI-Initiator, aber wenn dieser nicht enthalten ist, kann er problemlos installiert werden.

```
[root@host1 /]# yum install -y iscsi-initiator-utils
Loaded plugins: langpacks, product-id, search-disabled-repos,
subscription-
              : manager
Resolving Dependencies
--> Running transaction check
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.el7 will be
updated
--> Processing Dependency: iscsi-initiator-utils = 6.2.0.873-32.el7 for
package: iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7 will be
an update
--> Running transaction check
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.el7 will
be updated
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
```

```
will be an update
--> Finished Dependency Resolution
Dependencies Resolved
================================================================================
===
 Package                         Arch   Version              Repository
Size
================================================================================
===
Updating:
 iscsi-initiator-utils           x86_64 6.2.0.873-32.0.2.el7 ol7_latest 416
k
Updating for dependencies:
 iscsi-initiator-utils-iscsiuio x86_64 6.2.0.873-32.0.2.el7 ol7_latest  84
k
Transaction Summary
================================================================================
===
Upgrade  1 Package (+1 Dependent package)
Total download size: 501 k
Downloading packages:
No Presto metadata available for ol7_latest
(1/2): iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_6 | 416 kB   00:00
(2/2): iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2. |  84 kB   00:00
--------------------------------------------------------------------------------
---
Total                                           2.8 MB/s | 501 kB
00:00Cluster01
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Updating   : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
1/4
  Updating   : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
2/4
  Cleanup    : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Cleanup    : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
rhel-7-server-eus-rpms/7Server/x86_64/productid       | 1.7 kB     00:00
rhel-7-server-rpms/7Server/x86_64/productid           | 1.7 kB     00:00
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
1/4
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
2/4
```

```
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
Updated:
  iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7
Dependency Updated:
  iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
Complete!
[root@host1 /]#
```

Identifizieren Sie den iSCSI-Initiatornamen

Während der Installation wird ein eindeutiger iSCSI-Initiatorname generiert. Unter Linux befindet sie sich im `/etc/iscsi/initiatorname.iscsi` Datei: Dieser Name dient zur Identifizierung des Hosts auf dem IP-SAN.

```
[root@host1 /]# cat /etc/iscsi/initiatorname.iscsi
InitiatorName=iqn.1992-05.com.redhat:497bd66ca0
```

Erstellen Sie eine neue Initiatorgruppe

Eine Initiatorgruppe (Initiatorgruppe) ist Teil der ONTAP LUN-Masking-Architektur. Auf eine neu erstellte LUN kann nur dann zugegriffen werden, wenn einem Host der erste Zugriff gewährt wurde. Hierzu wird eine Initiatorgruppe erstellt, die entweder die FC-WWNs oder die iSCSI-Initiatornamen enthält, die Zugriff erfordern.

In diesem Beispiel wird eine Initiatorgruppe erstellt, die den iSCSI-Initiator des Linux Hosts enthält.

```
Cluster01::*> igroup create -igroup linuxiscsi -protocol iscsi -ostype
linux -initiator iqn.1994-05.com.redhat:497bd66ca0
```

Fahren Sie die Umgebung herunter

Vor dem Ändern des LUN-Protokolls müssen die LUNs vollständig stillgelegt werden. Jede Datenbank auf einer der zu konvertierenden LUNs muss heruntergefahren, die File-Systeme deaktiviert und die Volume-Gruppen deaktiviert werden. Wenn ASM verwendet wird, stellen Sie sicher, dass die ASM-Laufwerksgruppe getrennt ist und fahren Sie alle Netzdienste herunter.

LUN-Zuordnungen zum FC-Netzwerk aufheben

Nachdem die LUNs vollständig stillgelegt sind, entfernen Sie die Zuordnungen von der ursprünglichen FC-Initiatorgruppe.

```
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
```

LUN-Zuordnung zum IP-Netzwerk neu

Gewähren Sie der neuen iSCSI-basierten Initiatorgruppe Zugriff auf jede LUN.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxiscsi
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxiscsi
Cluster01::*>
```

ISCSI-Ziele erkennen

Die iSCSI-Erkennung besteht aus zwei Phasen. Zum einen werden die Ziele ermittelt. Dies ist nicht dasselbe wie beim Erkennen einer LUN. Der `iscsiadm` Der unten abgebildete Befehl prüft die vom angegebene Portalgruppe `-p argument` Und speichert eine Liste aller IP-Adressen und Ports, die iSCSI-Dienste anbieten. In diesem Fall gibt es vier IP-Adressen, die iSCSI-Dienste auf dem Standardport 3260 haben.

> Dieser Befehl kann mehrere Minuten dauern, wenn eine der Ziel-IP-Adressen nicht erreicht werden kann.

```
[root@host1 ~]# iscsiadm -m discovery -t st -p fas8060-iscsi-public1
10.63.147.197:3260,1033 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
10.63.147.198:3260,1034 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.203:3260,1030 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.202:3260,1029 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
```

ISCSI-LUNs erkennen

Nachdem die iSCSI-Ziele erkannt wurden, starten Sie den iSCSI-Dienst neu, um die verfügbaren iSCSI-LUNs zu ermitteln und zugehörige Geräte wie Multipath- oder ASMlib-Geräte zu erstellen.

```
[root@host1 ~]# service iscsi restart
Redirecting to /bin/systemctl restart  iscsi.service
```

Starten Sie die Umgebung neu

Starten Sie die Umgebung neu, indem Sie Volume-Gruppen erneut aktivieren, Dateisysteme neu mounten, RAC-Dienste neu starten usw. Als Vorsichtsmaßnahme empfiehlt NetApp, den Server nach Abschluss des Konvertierungsprozesses neu zu starten, um sicherzustellen, dass alle Konfigurationsdateien korrekt sind und alle veralteten Geräte entfernt werden.

Achtung: Bevor Sie einen Host neu starten, stellen Sie sicher, dass alle Einträge in sind `/etc/fstab` Diese Referenz migrierte SAN-Ressourcen werden kommentiert. Wenn dieser Schritt nicht durchgeführt wird und Probleme mit dem LUN-Zugriff auftreten, kann es zu einem Betriebssystem kommen, das nicht gebootet wird. Dieses Problem beschädigt die Daten nicht. Es kann jedoch sehr unbequem sein, in den Rettungsmodus oder einen ähnlichen Modus zu starten und zu korrigieren `/etc/fstab` Damit das Betriebssystem gestartet werden kann, um die Fehlerbehebung zu ermöglichen.

Beispielskripts

Die vorgestellten Skripte werden als Beispiele für das Skript verschiedener Betriebssystem- und Datenbankaufgaben bereitgestellt. Sie werden wie sie sind geliefert. Wenn für eine bestimmte Vorgehensweise Support erforderlich ist, wenden Sie sich an NetApp oder einen NetApp Reseller.

Datenbank wird heruntergefahren

Das folgende Perl-Skript nimmt ein einziges Argument der Oracle SID und fährt eine Datenbank herunter. Sie kann als Oracle-Benutzer oder als root ausgeführt werden.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
77 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF4
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
`;};
print @out;
if ("@out" =~ /ORACLE instance shut down/) {
print "$oraclesid shut down\n";
exit 0;}
elsif ("@out" =~ /Connected to an idle instance/) {
print "$oraclesid already shut down\n";
exit 0;}
else {
print "$oraclesid failed to shut down\n";
exit 1;}
```

Starten der Datenbank

Das folgende Perl-Skript nimmt ein einziges Argument der Oracle SID und fährt eine Datenbank herunter. Sie kann als Oracle-Benutzer oder als root ausgeführt werden.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
'
`;}
else {
@out=`. oraenv << EOF3
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
`;};
print @out;
if ("@out" =~ /Database opened/) {
print "$oraclesid started\n";
exit 0;}
elsif ("@out" =~ /cannot start already-running ORACLE/) {
print "$oraclesid already started\n";
exit 1;}
else {
78 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
print "$oraclesid failed to start\n";
exit 1;}
```

Konvertieren Sie das Dateisystem in schreibgeschützt

Das folgende Skript nimmt ein Dateisystemargument an und versucht, es als schreibgeschützt zu entfernen und wieder zu mounten. Dies ist bei Migrationsprozessen sinnvoll, bei denen ein Dateisystem für die Datenreplikation verfügbar gehalten werden muss und dennoch vor versehentlichen Schäden geschützt werden muss.

```
#! /usr/bin/perl
use strict;
#use warnings;
my $filesystem=$ARGV[0];
my @out=`umount '$filesystem'`;
if ($? == 0) {
 print "$filesystem unmounted\n";
 @out = `mount -o ro '$filesystem'`;
 if ($? == 0) {
 print "$filesystem mounted read-only\n";
 exit 0;}}
else {
 print "Unable to unmount $filesystem\n";
 exit 1;}
print @out;
```

Ersetzen Sie das Dateisystem

Das folgende Skriptbeispiel wird verwendet, um ein Dateisystem durch ein anderes zu ersetzen. Da die Datei `/etc/fstab `bearbeitet wird, muss sie als root ausgeführt werden. Es akzeptiert ein einzelnes kommagetrenntes Argument des alten und des neuen Dateisystems.

1. Führen Sie zum Ersetzen des Dateisystems das folgende Skript aus:

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oldfs;
my $newfs;
my @oldfstab;
my @newfstab;
my $source;
my $mountpoint;
my $leftover;
my $oldfstabentry='';
my $newfstabentry='';
my $migratedfstabentry='';
($oldfs, $newfs) = split (',',$ARGV[0]);
open(my $filehandle, '<', '/etc/fstab') or die "Could not open
/etc/fstab\n";
while (my $line = <$filehandle>) {
 chomp $line;
 ($source, $mountpoint, $leftover) = split(/[ , ]/,$line, 3);
 if ($mountpoint eq $oldfs) {
 $oldfstabentry = "#Removed by swap script $source $oldfs $leftover";}
```

```
 elsif ($mountpoint eq $newfs) {
 $newfstabentry = "#Removed by swap script $source $newfs $leftover";
 $migratedfstabentry = "$source $oldfs $leftover";}
 else {
 push (@newfstab, "$line\n")}}
79 Migration of Oracle Databases to NetApp Storage Systems © 2021
NetApp, Inc. All rights reserved
push (@newfstab, "$oldfstabentry\n");
push (@newfstab, "$newfstabentry\n");
push (@newfstab, "$migratedfstabentry\n");
close($filehandle);
if ($oldfstabentry eq ''){
 die "Could not find $oldfs in /etc/fstab\n";}
if ($newfstabentry eq ''){
 die "Could not find $newfs in /etc/fstab\n";}
my @out=`umount '$newfs'`;
if ($? == 0) {
 print "$newfs unmounted\n";}
else {
 print "Unable to unmount $newfs\n";
 exit 1;}
@out=`umount '$oldfs'`;
if ($? == 0) {
 print "$oldfs unmounted\n";}
else {
 print "Unable to unmount $oldfs\n";
 exit 1;}
system("cp /etc/fstab /etc/fstab.bak");
open ($filehandle, ">", '/etc/fstab') or die "Could not open /etc/fstab
for writing\n";
for my $line (@newfstab) {
 print $filehandle $line;}
close($filehandle);
@out=`mount '$oldfs'`;
if ($? == 0) {
 print "Mounted updated $oldfs\n";
 exit 0;}
else{
 print "Unable to mount updated $oldfs\n";
 exit 1;}
exit 0;
```

Nehmen Sie als Beispiel für die Verwendung dieses Skripts an, dass die Daten in enthalten sind `/oradata` Wird auf migriert `/neworadata` Und `/logs` Wird auf migriert `/newlogs`. Eine der einfachsten Methoden, um diese Aufgabe durchzuführen, besteht darin, das neue Gerät mit einem einfachen Dateikopiervorgang wieder in den ursprünglichen Bereitstellungspunkt zu verschieben.

2. Gehen Sie davon aus, dass die alten und die neuen Dateisysteme im vorhanden sind `/etc/fstab` Datei wie folgt:

```
cluster01:/vol_oradata /oradata nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
cluster01:/vol_logs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /newlogs nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
```

3. Wenn dieses Skript ausgeführt wird, wird das aktuelle Dateisystem abgehängt und durch das neue ersetzt:

```
[root@jfsc3 scripts]# ./swap.fs.pl /oradata,/neworadata
/neworadata unmounted
/oradata unmounted
Mounted updated /oradata
[root@jfsc3 scripts]# ./swap.fs.pl /logs,/newlogs
/newlogs unmounted
/logs unmounted
Mounted updated /logs
```

4. Das Skript aktualisiert auch die `/etc/fstab` Entsprechende Datei erstellen. Im hier gezeigten Beispiel sind folgende Änderungen enthalten:

```
#Removed by swap script cluster01:/vol_oradata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_logs /logs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_newlogs /newlogs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0
0
```

Automatisierte Datenbankmigration

Dieses Beispiel zeigt, wie Skripts zum Herunterfahren, Starten und Ersetzen von Dateisystemen genutzt werden können, um eine Migration vollständig zu automatisieren.

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my @oldfs;
my @newfs;
my $x=1;
while ($x < scalar(@ARGV)) {
   ($oldfs[$x-1], $newfs[$x-1]) = split (',',$ARGV[$x]);
   $x+=1;}
my @out=`./dbshut.pl '$oraclesid'`;
print @out;
if ($? ne 0) {
   print "Failed to shut down database\n";
   exit 0;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`./mk.fs.readonly.pl '$oldfs[$x]'`;
   if ($? ne 0) {
      print "Failed to make filesystem $oldfs[$x] readonly\n";
      exit 0;}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`rsync -rlpogt --stats --progress --exclude='.snapshot'
'$oldfs[$x]/' '/$newfs[$x]/'`;
   print @out;
   if ($? ne 0) {
      print "Failed to copy filesystem $oldfs[$x] to $newfs[$x]\n";
      exit 0;}
   else {
      print "Succesfully replicated filesystem $oldfs[$x] to
$newfs[$x]\n";}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   print "swap $x $oldfs[$x] $newfs[$x]\n";
   my @out=`./swap.fs.pl '$oldfs[$x],$newfs[$x]'`;
   print @out;
   if ($? ne 0) {
      print "Failed to swap filesystem $oldfs[$x] for $newfs[$x]\n";
      exit 1;}
   else {
      print "Swapped filesystem $oldfs[$x] for $newfs[$x]\n";}
   $x+=1;}
my @out=`./dbstart.pl '$oraclesid'`;
```

```
print @out;
```

Dateispeicherorte anzeigen

Dieses Skript sammelt eine Reihe wichtiger Datenbankparameter und druckt sie in einem leicht lesbaren Format aus. Dieses Skript kann bei der Überprüfung von Datenlayouts nützlich sein. Darüber hinaus kann das Skript für andere Zwecke geändert werden.

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;};
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print "$oraclesid datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
```

```
@out=dosql('select member from v\\\\\$logfile;');
print "$oraclesid redo logs:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name from v\\\\\$tempfile;');
print "$oraclesid temp datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('show parameter spfile;');
print "$oraclesid spfile\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name||\' \'||value from v\\\\\$parameter where
isdefault=\'FALSE\';');
print "$oraclesid key parameters\n";
for $line (@out) {
        chomp($line);
        if ($line =~ /control_files/) {print "$line\n";}
        if ($line =~ /db_create/) {print "$line\n";}
        if ($line =~ /db_file_name_convert/) {print "$line\n";}
        if ($line =~ /log_archive_dest/) {print "$line\n";}}
        if ($line =~ /log_file_name_convert/) {print "$line\n";}
        if ($line =~ /pdb_file_name_convert/) {print "$line\n";}
        if ($line =~ /spfile/) {print "$line\n";}
print "\n";
```

Bereinigung der ASM-Migration

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
```

```
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;}
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print @out;
print "shutdown immediate;\n";
print "startup mount;\n";
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "host mv $line $1$newname\n";}}
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "alter database rename file '$line' to
'$1$newname';\n";}}
```

```
print "alter database open;\n";
print "\n";
```

Namenskonvertierung von ASM in Dateisystem

```
set serveroutput on;
set wrap off;
declare
    cursor df is select file#, name from v$datafile;
    cursor tf is select file#, name from v$tempfile;
    cursor lf is select member from v$logfile;
    firstline boolean := true;
begin
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('Parameters for log file conversion:');
    dbms_output.put_line(CHR(13));
    dbms_output.put('*.log_file_name_convert = ');
    for lfrec in lf loop
        if (firstline = true) then
            dbms_output.put('''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*/','') || '''');
        else
            dbms_output.put(',''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*/','') || '''');
        end if;
        firstline:=false;
    end loop;
    dbms_output.put_line(CHR(13));
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('rman duplication script:');
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('run');
    dbms_output.put_line('{');
    for dfrec in df loop
        dbms_output.put_line('set newname for datafile ' ||
            dfrec.file# || ' to ''' || dfrec.name ||''';');
    end loop;
    for tfrec in tf loop
        dbms_output.put_line('set newname for tempfile ' ||
            tfrec.file# || ' to ''' || tfrec.name ||''';');
    end loop;
    dbms_output.put_line('duplicate target database for standby backup
location INSERT_PATH_HERE;');
    dbms_output.put_line('}');
end;
/
```

Wiedergabe von Protokollen in der Datenbank

Dieses Skript akzeptiert ein einzelnes Argument einer Oracle SID für eine Datenbank, die sich im Mount-Modus befindet, und versucht, alle derzeit verfügbaren Archivprotokolle wiederzugeben.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
84 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
`;
}
print @out;
```

Wiedergabe von Protokollen in der Standby-Datenbank

Dieses Skript ist identisch mit dem vorhergehenden Skript, außer dass es für eine Standby-Datenbank konzipiert ist.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
`;
}
print @out;
```

Zusätzliche Anmerkungen

Performance-Optimierung und Benchmarking

Das genaue Testen der Datenbank-Storage-Performance ist dabei ein extrem kompliziertes Thema. Sie müssen die folgenden Probleme verstehen:

- IOPS und Durchsatz
- Der Unterschied zwischen Vorder- und Hintergrund-I/O-Vorgängen
- Auswirkungen der Latenz auf die Datenbank
- Zahlreiche Betriebssystem- und Netzwerkeinstellungen, die ebenfalls die Storage-Performance beeinträchtigen

Darüber hinaus müssen Aufgaben außerhalb von Storage-Datenbanken berücksichtigt werden. An diesem Punkt kann die Optimierung der Storage-Performance keine nützlichen Vorteile ergeben, da die Storage-Performance keinen einschränkenden Faktor mehr für die Performance darstellt.

Da sich die meisten Datenbankkunden nun für All-Flash-Arrays entscheiden, sind weitere Überlegungen anzustellen. Betrachten Sie beispielsweise Performance-Tests auf einem AFF A900 System mit zwei Nodes:

- Mit einem Lese-/Schreib-Verhältnis von 80/20 können zwei A900 Nodes über 1 Mio. zufällige Datenbank-IOPS liefern, bevor die Latenz sogar die 150µs-Marke überschreitet. Dies geht weit über die aktuellen Performance-Anforderungen der meisten Datenbanken hinaus, sodass sich die erwartete Verbesserung nur schwer vorhersagen lässt. Storage würde zu einem großen Teil als Engpass gelöscht werden.
- Die Netzwerkbandbreite ist eine immer häufiger auftretende Ursache für Leistungseinschränkungen. Lösungen mit rotierenden Festplatten sind beispielsweise häufig Engpässe in der Datenbank-Performance, da die I/O-Latenz sehr hoch ist. Wenn Latenzbeschränkungen von einem All-Flash-Array beseitigt werden, verschiebt sich die Barriere häufig in das Netzwerk. Dies ist insbesondere bei virtualisierten Umgebungen und Blade-Systemen so bemerkenswert, dass sich die tatsächliche Netzwerkverbindung nur schwer visualisieren lässt. Das kann Performance-Tests erschweren, wenn das Storage-System selbst aufgrund von Bandbreiteneinschränkungen nicht vollständig ausgelastet werden kann.
- Der Vergleich der Performance eines All-Flash-Arrays mit einem Array mit rotierenden Festplatten ist im Allgemeinen aufgrund der deutlich verbesserten Latenz von All-Flash-Arrays nicht möglich. Die Testergebnisse sind in der Regel nicht aussagekräftig.
- Der Vergleich der IOPS-Spitzenperformance mit einem All-Flash-Array ist häufig kein nützlicher Test, da Datenbanken nicht durch Storage-I/O eingeschränkt werden Angenommen, ein Array unterstützt 500.000 zufällige IOPS, ein anderes dagegen 300.000. Der Unterschied ist in der Praxis irrelevant, wenn eine Datenbank 99 % ihrer Zeit für die CPU-Verarbeitung aufwendet. Die Workloads schöpfen dabei niemals alle Kapazitäten des Storage-Arrays aus. Demgegenüber sind IOPS-Spitzenfunktionen bei einer Konsolidierungsplattform durchaus von großer Bedeutung, bei der das Storage-Array voraussichtlich auf seine Spitzenfunktionen ausgelastet ist.
- Berücksichtigen Sie bei jedem Storage-Test sowohl Latenz als auch IOPS. Viele Storage Arrays auf dem Markt bieten angeblich ein äußerst extremes IOPS-Niveau, doch aufgrund der Latenz sind diese IOPS in einem solchen Maß nutzlos. Ein typisches Ziel mit All-Flash-Arrays ist die Marke von 1 ms. Ein besserer Testansatz ist nicht die Messung der maximal möglichen IOPS, sondern die Ermittlung der IOPS-Anzahl, die ein Storage-Array verarbeiten kann, bevor die durchschnittliche Latenz größer als 1 ms ist.

Oracle Automatic Workload Repository und Benchmarking

Der Goldstandard für die Performance-Vergleiche mit Oracle ist ein Oracle Automatic Workload Repository (AWR) Bericht.

Es gibt mehrere Typen von AWR-Berichten. Aus Sicht des Speicherpunkts ein Bericht, der durch Ausführen des generiert wird `awrrpt.sql` Der Befehl ist der umfassendste und wertvollste, da er auf eine bestimmte Datenbankinstanz abzielt und einige detaillierte Histogramme enthält, die Speicher-I/O-Ereignisse basierend auf Latenz aufteilen.

Um zwei Performance-Arrays zu vergleichen, wird idealerweise derselbe Workload auf jedem Array ausgeführt und ein AWR-Bericht erstellt, der genau auf den Workload abzielt. Bei einer sehr langen Arbeitslast kann ein einzelner AWR-Bericht mit einer verstrichenen Zeit, die die Start- und Stoppzeit umfasst, verwendet werden. Es ist jedoch vorzuziehen, die AWR-Daten als mehrere Berichte auszuteilen. Wenn beispielsweise ein Batch-Job von Mitternacht bis 6 Uhr ausgeführt wurde, erstellen Sie eine Reihe einstündiger AWR-Berichte von Mitternacht bis 1 Uhr, von 1 bis 2 Uhr usw.

In anderen Fällen sollte eine sehr kurze Abfrage optimiert werden. Die beste Option ist ein AWR-Bericht, der auf einem AWR-Snapshot basiert, der beim Start der Abfrage erstellt wurde, und ein zweiter AWR-Snapshot, der beim Ende der Abfrage erstellt wurde. Der Datenbankserver sollte ansonsten ruhig sein, um die Hintergrundaktivität zu minimieren, die die Aktivität der analysierten Abfrage verdunkeln würde.

Wenn AWR-Berichte nicht verfügbar sind, sind Oracle Statspack-Berichte eine gute Alternative. Sie enthalten die meisten der gleichen I/O-Statistiken wie ein AWR-Bericht.

Oracle AWR und Fehlerbehebung

Ein AWR-Bericht ist auch das wichtigste Werkzeug zur Analyse eines Leistungsproblems.

Wie bei Benchmarking muss auch bei der Performance-Fehlerbehebung ein bestimmter Workload genau gemessen werden. Wenn möglich, geben Sie AWR-Daten ein, wenn Sie dem NetApp Support Center ein Performance-Problem melden oder wenn Sie mit einem NetApp oder einem Partner Account Team an einer neuen Lösung arbeiten.

Beachten Sie bei der Bereitstellung von AWR-Daten die folgenden Anforderungen:

- Führen Sie die aus `awrrpt.sql` Befehl zum Generieren des Berichts. Die Ausgabe kann entweder Text oder HTML sein.
- Wenn Oracle Real Application Clusters (RACs) verwendet werden, erstellen Sie AWR-Berichte für jede Instanz im Cluster.
- Ziel der Zeitpunkt, zu dem das Problem aufgetreten ist. Die maximal zulässige verstrichene Zeit eines AWR-Berichts beträgt in der Regel eine Stunde. Wenn ein Problem mehrere Stunden andauert oder einen mehrstündigen Vorgang wie einen Batch-Job umfasst, stellen Sie mehrere einstündige AWR-Berichte bereit, die den gesamten zu analysierenden Zeitraum abdecken.
- Wenn möglich, stellen Sie das AWR-Snapshot-Intervall auf 15 Minuten ein. Diese Einstellung ermöglicht eine detailliertere Analyse. Dies erfordert auch zusätzliche Ausführungen von `awrrpt.sql` Um einen Bericht für jedes 15-Minuten-Intervall bereitzustellen.
- Wenn es sich bei dem Problem um eine sehr kurze laufende Abfrage handelt, geben Sie einen AWR-Bericht an, der auf einem AWR-Snapshot basiert, der beim Start des Vorgangs erstellt wurde, und einen zweiten AWR-Snapshot, der nach Beendigung des Vorgangs erstellt wurde. Der Datenbankserver sollte ansonsten ruhig sein, um die Hintergrundaktivität zu minimieren, die die Aktivität des analysierten Vorgangs verdunkeln würde.
- Wenn ein Leistungsproblem zu bestimmten Zeiten gemeldet wird, aber nicht zu anderen, liefern Sie zusätzliche AWR-Daten, die eine gute Leistung zum Vergleich zeigen.

Kalibrieren_io

Der `calibrate_io` Der Befehl sollte niemals zum Testen, Vergleichen oder Vergleichen von Storage-Systemen verwendet werden. Wie in der Oracle-Dokumentation beschrieben, werden mit diesem Verfahren die I/O-Funktionen des Speichers kalibriert.

Kalibrierung ist nicht dasselbe wie Benchmarking. Mit diesem Befehl können Sie I/O-Vorgänge ausgeben, um Datenbankvorgänge zu kalibrieren und ihre Effizienz zu verbessern, indem Sie die I/O-Ausgabe für den Host optimieren. Da der Typ der I/O, die vom ausgeführt wird `calibrate_io` Der Betrieb entspricht nicht der tatsächlichen I/O von Datenbankbenutzern, die Ergebnisse sind nicht vorhersehbar und häufig nicht einmal reproduzierbar.

SLOB2

SLOB2, der Silly Little Oracle Benchmark, ist zum bevorzugten Tool für die Bewertung der Datenbank-Performance geworden. Es wurde von Kevin Closson entwickelt und ist verfügbar unter "https://kevinclosson.net/slob/". Die Installation und Konfiguration dauert nur wenige Minuten und mithilfe einer echten Oracle-Datenbank lassen sich I/O-Muster auf einem benutzerdefinierbaren Tablespace generieren. Es

ist eine der wenigen verfügbaren Testoptionen, die die Auslastung eines All-Flash-Arrays mit I/O-Vorgängen ermöglichen Er eignet sich auch zur Generierung von deutlich niedrigeren I/O-Werten, um Storage-Workloads zu simulieren, die zwar niedrige IOPS, aber latenzempfindlich sind.

Wechselbank

SwingBench kann zum Testen der Datenbank-Performance nützlich sein, aber es ist extrem schwierig, SwingBench auf eine Art und Weise zu verwenden, die den Storage belastet. Bei NetApp gab es noch keine Tests von Swingbench, die genug I/O ergaben, um auf jedem AFF Array eine erhebliche Belastung zu sein. In begrenzten Fällen kann der Order Entry Test (OET) verwendet werden, um die Storage-Systeme unter Latenzsicht zu bewerten. Dies kann in Situationen nützlich sein, in denen eine Datenbank eine bekannte Latenzabhängigkeit für bestimmte Abfragen hat. Achten Sie unbedingt darauf, dass Host und Netzwerk ordnungsgemäß konfiguriert sind, um die Latenzpotenziale eines All-Flash-Arrays auszuschöpfen.

HammerDB

HammerDB ist ein Datenbank-Test-Tool, das unter anderem TPC-C- und TPC-H-Benchmarks simuliert. Es kann eine Menge Zeit dauern, bis ein ausreichend großer Datensatz für die ordnungsgemäße Ausführung eines Tests erstellt wurde. Er kann aber ein effektives Tool zur Performance-Evaluierung für OLTP- und Data Warehouse-Applikationen sein.

Orion

Das Oracle Orion Tool wurde häufig mit Oracle 9 verwendet, wurde jedoch nicht gewartet, um die Kompatibilität mit Änderungen in verschiedenen Host-Betriebssystemen zu gewährleisten. Er wird aufgrund der Inkompatibilitäten mit der Betriebssystem- und Storage-Konfiguration selten mit Oracle 10 oder Oracle 11 verwendet.

Oracle hat das Tool neu geschrieben und es wird standardmäßig mit Oracle 12c installiert. Obwohl dieses Produkt verbessert wurde und viele der gleichen Aufrufe verwendet, die eine echte Oracle-Datenbank verwendet, verwendet es nicht genau den gleichen Codepfad oder das gleiche I/O-Verhalten, das von Oracle verwendet wird. Beispielsweise werden die meisten Oracle I/OS synchron ausgeführt, was bedeutet, dass die Datenbank angehalten wird, bis der I/O-Vorgang abgeschlossen ist, während der I/O-Vorgang im Vordergrund abgeschlossen ist. Eine einfache Überflutung eines Storage-Systems mit zufälligen I/OS ist keine Reproduktion von realen Oracle I/O und bietet keine direkte Methode, Storage Arrays zu vergleichen oder die Auswirkungen von Konfigurationsänderungen zu messen.

Dennoch gibt es einige Anwendungsfälle für Orion, wie z. B. die generelle Messung der maximal möglichen Performance einer bestimmten Host-Netzwerk-Storage-Konfiguration oder die Abmessung des Zustands eines Storage-Systems. Mit sorgfältigen Tests können nutzbare Orion Tests entwickelt werden, um Storage-Arrays zu vergleichen oder die Auswirkungen einer Konfigurationsänderung zu bewerten, sofern zu den Parametern IOPS, Durchsatz und Latenz gehören und versucht werden, einen realistischen Workload originalgetreu zu replizieren.

Veraltete NFSv3-Sperren

Wenn ein Oracle-Datenbankserver abstürzt, kann es beim Neustart zu Problemen mit veralteten NFS-Sperren kommen. Dieses Problem ist vermeidbar, indem Sie sorgfältig auf die Konfiguration der Namensauflösung auf dem Server achten.

Dieses Problem tritt auf, weil das Erstellen einer Sperre und das Löschen einer Sperre zwei leicht unterschiedliche Methoden der Namensauflösung verwenden. Es sind zwei Prozesse beteiligt: Der Network Lock Manager (NLM) und der NFS-Client. Der NLM verwendet `uname n` Um den Hostnamen zu ermitteln, während der `rpc.statd` Prozessanwendungen `gethostbyname()`. Diese Hostnamen müssen

übereinstimmen, damit das Betriebssystem veraltete Sperren ordnungsgemäß löschen kann. Beispielsweise sucht der Host nach Sperren, die Eigentum von sind `dbserver5`, Aber die Schlösser wurden vom Host als registriert `dbserver5.mydomain.org`. Wenn `gethostbyname()` Gibt nicht denselben Wert zurück wie `uname -a`, Dann ist der Sperrvorgang nicht erfolgreich.

Mit dem folgenden Beispielskript wird überprüft, ob die Namensauflösung vollständig konsistent ist:

```
#! /usr/bin/perl
$uname=`uname -n`;
chomp($uname);
($name, $aliases, $addrtype, $length, @addrs) = gethostbyname $uname;
print "uname -n yields: $uname\n";
print "gethostbyname yields: $name\n";
```

Wenn `gethostbyname` Stimmt nicht überein `uname`, Veraltete Sperren sind wahrscheinlich. Dieses Ergebnis zeigt beispielsweise ein potenzielles Problem auf:

```
uname -n yields: dbserver5
gethostbyname yields: dbserver5.mydomain.org
```

Die Lösung wird normalerweise durch Ändern der Reihenfolge gefunden, in der Hosts in angezeigt werden `/etc/hosts`. Nehmen wir beispielsweise an, dass die Hosts-Datei diesen Eintrag enthält:

```
10.156.110.201  dbserver5.mydomain.org dbserver5 loghost
```

Um dieses Problem zu beheben, ändern Sie die Reihenfolge, in der der vollständig qualifizierte Domänenname und der kurze Hostname angezeigt werden:

```
10.156.110.201  dbserver5 dbserver5.mydomain.org loghost
```

`gethostbyname()` Gibt nun den Short zurück `dbserver5` Host-Name, der mit der Ausgabe von übereinstimmt `uname`. Sperren werden somit nach einem Serverabsturz automatisch gelöscht.

Überprüfung der WAFL-Ausrichtung

Eine korrekte WAFL-Ausrichtung ist für eine gute Performance von entscheidender Bedeutung. Obwohl ONTAP Blöcke in 4-KB-Einheiten managt, bedeutet dies nicht, dass ONTAP alle Vorgänge in 4-KB-Einheiten ausführt. ONTAP unterstützt zwar Blockoperationen unterschiedlicher Größen, die zugrunde liegende Buchhaltung wird jedoch von WAFL in 4-KB-Einheiten gemanagt.

Der Begriff „Alignment“ bezieht sich darauf, wie Oracle I/O diesen 4-KB-Einheiten entspricht. Für eine optimale Performance ist ein Oracle 8-KB-Block auf zwei physischen 4-KB-WAFL-Blöcken eines Laufwerks erforderlich. Wenn ein Block durch 2 KB verrechnet wird, befindet sich dieser Block auf der Hälfte eines 4-KB-Blocks, einem separaten vollständigen 4-KB-Block und dann der Hälfte eines dritten 4-KB-Blocks. Diese Anordnung

führt zu Leistungseinbußen.

Bei NAS-File-Systemen ist die Ausrichtung nicht relevant. Oracle Datendateien werden am Anfang der Datei auf Basis der Größe des Oracle Blocks ausgerichtet. Daher sind Blockgrößen von 8 KB, 16 KB und 32 KB immer ausgerichtet. Alle Blockoperationen werden vom Anfang der Datei in Einheiten von 4 Kilobyte versetzt.

LUNs enthalten dagegen in der Regel zu Beginn eine Art Treiber-Header oder Filesystem-Metadaten, wodurch ein Offset erzeugt wird. Die Ausrichtung ist in modernen Betriebssystemen selten ein Problem, da diese Betriebssysteme für physische Laufwerke ausgelegt sind, die möglicherweise einen nativen 4-KB-Sektor verwenden, was außerdem die Ausrichtung der I/O an 4-KB-Grenzen erfordert, um eine optimale Performance zu erzielen.

Es gibt jedoch einige Ausnahmen. Eine Datenbank wurde möglicherweise von einem älteren Betriebssystem migriert, das nicht für 4 KB I/O optimiert wurde, oder ein Benutzerfehler während der Partitionserstellung hat möglicherweise zu einem Offset geführt, der sich nicht in Einheiten von 4 KB befindet.

Die folgenden Beispiele sind Linux-spezifisch, aber das Verfahren kann für jedes Betriebssystem angepasst werden.

Ausgerichtet

Das folgende Beispiel zeigt eine Ausrichtungsüberprüfung einer einzelnen LUN mit einer einzelnen Partition.

Erstellen Sie zunächst die Partition, die alle auf dem Laufwerk verfügbaren Partitionen verwendet.

```
[root@host0 iscsi]# fdisk /dev/sdb
Device contains neither a valid DOS partition table, nor Sun, SGI or OSF
disklabel
Building a new DOS disklabel with disk identifier 0xb97f94c1.
Changes will remain in memory only, until you decide to write them.
After that, of course, the previous content won't be recoverable.
The device presents a logical sector size that is smaller than
the physical sector size. Aligning to a physical sector (or optimal
I/O) size boundary is recommended, or performance may be impacted.
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-10240, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-10240, default 10240):
Using default value 10240
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
[root@host0 iscsi]#
```

Die Ausrichtung kann mathematisch mit folgendem Befehl überprüft werden:

```
[root@host0 iscsi]# fdisk -u -l /dev/sdb
Disk /dev/sdb: 10.7 GB, 10737418240 bytes
64 heads, 32 sectors/track, 10240 cylinders, total 20971520 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 4096 bytes
I/O size (minimum/optimal): 4096 bytes / 65536 bytes
Disk identifier: 0xb97f94c1
   Device Boot      Start         End      Blocks   Id  System
/dev/sdb1              32    20971519    10485744   83  Linux
```

Die Ausgabe zeigt an, dass die Einheiten 512 Byte betragen, und der Beginn der Partition ist 32 Einheiten. Dies sind insgesamt 32 x 512 = 16,834 Byte, was ein ganzes Vielfaches von 4-KB-WAFL-Blöcken ist. Diese Partition ist korrekt ausgerichtet.

Führen Sie die folgenden Schritte durch, um die korrekte Ausrichtung zu überprüfen:

1. Identifizieren Sie die UUID (Universally Unique Identifier) der LUN.

```
FAS8040SAP::> lun show -v /vol/jfs_luns/lun0
              Vserver Name: jfs
                  LUN UUID: ed95d953-1560-4f74-9006-85b352f58fcd
                    Mapped: mapped`                `
```

2. Geben Sie die Node-Shell auf dem ONTAP-Controller ein.

```
FAS8040SAP::> node run -node FAS8040SAP-02
Type 'exit' or 'Ctrl-D' to return to the CLI
FAS8040SAP-02> set advanced
set not found.  Type '?' for a list of commands
FAS8040SAP-02> priv set advanced
Warning: These advanced commands are potentially dangerous; use
         them only when directed to do so by NetApp
         personnel.
```

3. Starten Sie statistische Sammlungen auf der im ersten Schritt identifizierten Ziel-UUID.

```
FAS8040SAP-02*> stats start lun:ed95d953-1560-4f74-9006-85b352f58fcd
Stats identifier name is 'Ind0xffffff08b9536188'
FAS8040SAP-02*>
```

4. Führen Sie einige I/O-Vorgänge aus Es ist wichtig, die zu verwenden `iflag` Argument, um sicherzustellen, dass I/O synchron und nicht gepuffert ist.

Seien Sie sehr vorsichtig mit diesem Befehl. Umkehren der `if` Und `of` Argumente zerstören Daten.

```
[root@host0 iscsi]# dd if=/dev/sdb1 of=/dev/null iflag=dsync count=1000
bs=4096
1000+0 records in
1000+0 records out
4096000 bytes (4.1 MB) copied, 0.0186706 s, 219 MB/s
```

5. Stoppen Sie die Statistiken, und zeigen Sie das Alignment-Histogramm an. Alle I/O-Vorgänge sollten sich im befinden `.0` Bucket-Modul zur Angabe von I/O, die an einer 4-KB-Blockgrenze ausgerichtet ist

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff08b9536188
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:186%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
```

Falsch Ausgerichtet

Im folgenden Beispiel wird eine falsch ausgerichtete I/O angezeigt:

1. Erstellen Sie eine Partition, die nicht an einer 4-KB-Grenze ausgerichtet ist. Dies ist kein Standardverhalten auf modernen Betriebssystemen.

```
[root@host0 iscsi]# fdisk -u /dev/sdb
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (32-20971519, default 32): 33
Last sector, +sectors or +size{K,M,G} (33-20971519, default 20971519):
Using default value 20971519
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

2. Die Partition wurde mit einem 33-Sektor-Offset anstelle der Standardeinstellung 32 erstellt. Wiederholen Sie den in beschriebenen Vorgang "Ausgerichtet". Das Histogramm wird wie folgt angezeigt:

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:136%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_partial_blocks:31%
```

Die Fehlausrichtung ist klar. Die I/O fällt meist in das* *`.1` Bucket, der dem erwarteten Offset entspricht. Bei der Erstellung der Partition wurde sie 512 Byte weiter in das Gerät verschoben als der optimierte Standardwert, was bedeutet, dass das Histogramm durch 512 Byte versetzt wird.

Darüber hinaus der `read_partial_blocks` Die Statistik ist ein Wert ungleich Null, was bedeutet, dass I/O-Vorgänge ausgeführt wurden, die keinen gesamten 4-KB-Block aufgefüllt haben.

Wiederherstellungsprotokollierung

Die hier erläuterten Verfahren gelten für Datendateien. Oracle Redo- und Archivprotokolle weisen unterschiedliche I/O-Muster auf. Beispielsweise ist die Wiederherstellungsprotokollierung ein kreisförmiges Überschreiben einer einzelnen Datei. Wenn die standardmäßige 512-Byte-Blockgröße verwendet wird, sehen die Schreibstatistiken in etwa wie folgt aus:

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-4f74-
9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.0:12%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.1:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.3:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.4:13%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.5:6%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.6:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.7:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_partial_blocks:85%
```

Die I/O-Vorgänge werden auf alle Histogramm-Buckets verteilt, dies stellt jedoch keine Performance-Sorge dar. Extrem hohe Redo-Protokollierungsraten können jedoch von der Verwendung einer 4-KB-Blockgröße profitieren. In diesem Fall ist es wünschenswert, dass die LUNs für die Wiederherstellungsprotokollierung ordnungsgemäß ausgerichtet sind. Dies ist jedoch für eine gute Performance nicht so wichtig wie die Datendateiausrichtung.