Storage-Konfiguration auf AFF/FAS Systemen

Enterprise applications

NetApp
January 12, 2026

Inhalt

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.

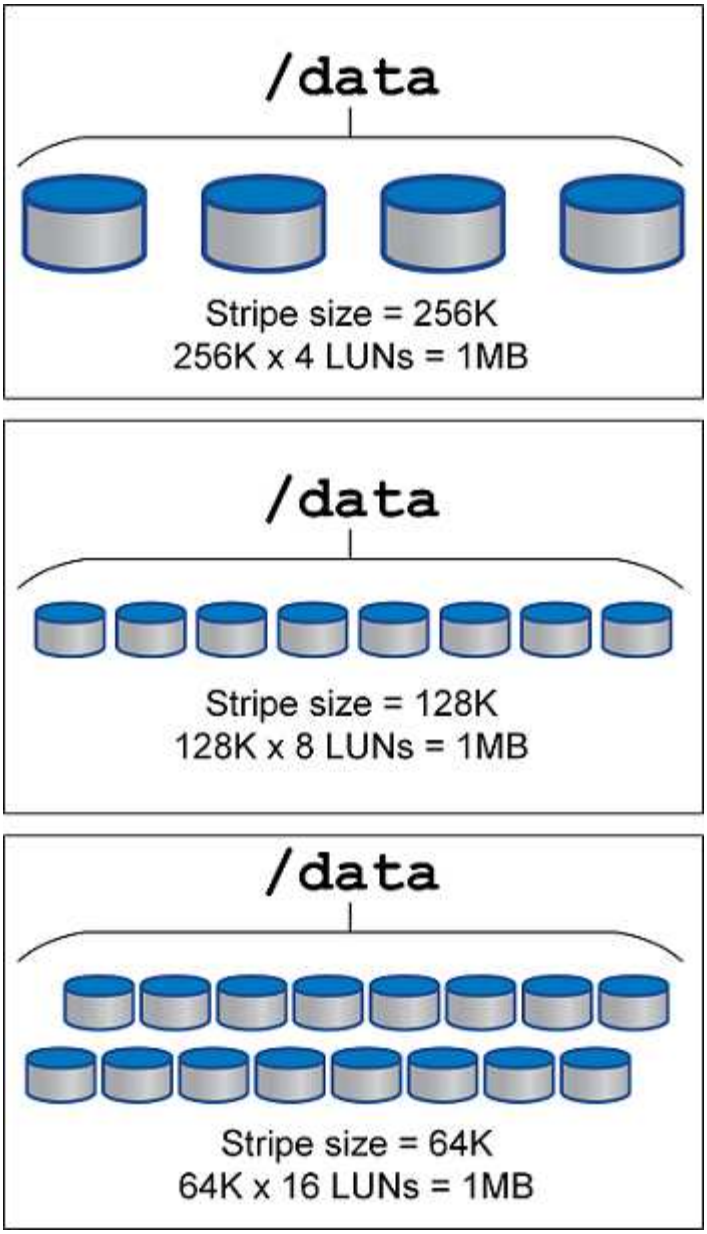
Das Striping mit rotierenden Laufwerken war schwieriger, da das I/O-Muster bereits im Vorfeld bekannt sein musste. Wenn das Striping nicht richtig auf die wahren I/O-Muster abgestimmt wurde, können Striping-Konfigurationen die Performance beeinträchtigen. Bei Oracle Datenbanken und insbesondere bei All-Flash-Konfigurationen 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.