Database Oracle

Enterprise applications

NetApp
January 12, 2026

Sommario

Database Oracle

Database Oracle su ONTAP

ONTAP è progettato per i database Oracle. Per decenni, ONTAP è stato ottimizzato per le esigenze uniche di i/o dei database relazionali e sono state create più funzionalità di ONTAP appositamente per soddisfare le esigenze dei database Oracle e persino su richiesta della stessa Oracle Inc.

> Questa documentazione sostituisce i report tecnici precedentemente pubblicati *TR-3633: Database Oracle su ONTAP; TR-4591: Protezione dei dati Oracle: Backup, recovery, replica; TR-4592: Oracle su MetroCluster; e TR-4534: Migrazione dei database Oracle su sistemi di storage NetApp*

Oltre ai numerosi modi possibili in cui ONTAP apporta valore all'ambiente di database, esiste anche una vasta gamma di requisiti utente, incluse le dimensioni del database, i requisiti di performance e le esigenze di protezione dei dati. Le distribuzioni note di storage NetApp includono tutto, da un ambiente virtualizzato di circa 6.000 database in esecuzione su VMware ESX a un data warehouse a singola istanza, di dimensioni attuali pari a 996TB TB e in crescita. Di conseguenza, sono disponibili alcune Best practice chiare per la configurazione di un database Oracle su storage NetApp.

I requisiti per l'utilizzo di un database Oracle su storage NetApp vengono risolti in due modi. In primo luogo, quando esiste una buona pratica chiara, essa verrà richiamata in modo specifico. A un livello generale, verranno spiegate molte considerazioni di progettazione che i progettisti delle soluzioni di storage Oracle devono affrontare in base ai loro specifici requisiti di business.

Configurazione ONTAP sui sistemi AFF/ FAS

RAID

RAID si riferisce all'utilizzo della ridondanza per proteggere i dati dalla perdita di un'unità.

Occasionalmente sorgono domande riguardanti i livelli RAID nella configurazione dello storage NetApp utilizzato per i database Oracle e altre applicazioni aziendali. Molte Best practice Oracle precedenti relative alla configurazione degli array di storage contengono avvisi sull'utilizzo del mirroring RAID e/o sull'eliminazione di determinati tipi di RAID. Sebbene sollevino punti validi, questi sorgenti non si applicano a RAID 4 e alle tecnologie NetApp RAID DP e RAID-TEC utilizzate in ONTAP.

RAID 4, RAID 5, RAID 6, RAID DP e RAID-TEC utilizzano tutti la parità per garantire che il guasto al disco non determini una perdita di dati. Queste opzioni RAID offrono un utilizzo dello storage migliore rispetto al mirroring, ma la maggior parte delle implementazioni RAID presenta uno svantaggio che influisce sulle operazioni di scrittura. Il completamento di un'operazione di scrittura su altre implementazioni RAID potrebbe richiedere letture di più unità per rigenerare i dati di parità, un processo comunemente chiamato penalizzazione RAID.

ONTAP, tuttavia, non subisce questa penalizzazione del RAID. Ciò è dovuto all'integrazione di NetApp WAFL (Write Anywhere file Layout) con il livello RAID. Le operazioni di scrittura vengono unite nella RAM e preparate come uno stripe RAID completo, inclusa la generazione della parità. ONTAP non ha bisogno di eseguire una lettura per completare una scrittura, il che significa che ONTAP e WAFL evitare la penalizzazione RAID. Le performance per le operazioni critiche in termini di latenza, come il logging di redo, vengono mantenute e le scritture random dei file di dati non comportano penalizzazioni RAID dovute alla necessità di rigenerare la

parità.

Per quanto riguarda l'affidabilità statistica, anche RAID DP offre una protezione migliore rispetto al mirroring RAID. Il problema principale è la richiesta fatta sui dischi durante una ricostruzione del RAID. Con un set RAID con mirroring, il rischio di perdita di dati causata da un guasto al disco e durante la ricostruzione nel partner nel set RAID è molto maggiore del rischio di un guasto a tre dischi in un set RAID DP.

Gestione della capacità

La gestione di un database o di un'altra applicazione aziendale con storage aziendale prevedibile, gestibile e ad alte prestazioni richiede spazio libero sulle unità per la gestione di dati e metadati. La quantità di spazio libero richiesta dipende dal tipo di unità utilizzata e dai processi aziendali.

Lo spazio libero viene definito come lo spazio non utilizzato per i dati effettivi e include lo spazio non allocato dell'aggregato e lo spazio non utilizzato all'interno dei volumi costituenti. È importante prendere in considerazione anche il thin provisioning. Ad esempio, un volume potrebbe contenere un LUN da 1TB GB, di cui solo il 50% viene utilizzato dai dati reali. In un ambiente con thin provisioning, questo sembra consumare correttamente 500GB GB di spazio. Tuttavia, in un ambiente con provisioning completo, la capacità completa di 1TB TB sembra essere in uso. I 500GB GB di spazio non allocato sono nascosti. Questo spazio non è utilizzato dai dati effettivi e deve quindi essere incluso nel calcolo dello spazio libero totale.

Di seguito sono riportate le raccomandazioni NetApp per i sistemi storage utilizzati per le applicazioni aziendali:

Aggregati SSD, inclusi i sistemi AFF

NetApp consiglia almeno il 10% di spazio libero. Ciò comprende tutto lo spazio inutilizzato, compreso lo spazio libero all'interno dell'aggregato o di un volume ed eventuale spazio libero allocato a causa dell'utilizzo del provisioning completo, ma non utilizzato dai dati effettivi. Lo spazio logico non è importante, la domanda è quanto spazio fisico libero effettivo è disponibile per lo storage dei dati.

Il consiglio di liberare il 10% dello spazio è molto conservativo. Gli aggregati SSD possono supportare i carichi di lavoro a livelli di utilizzo ancora più elevati senza influire sulle performance. Tuttavia, con l'aumento dell'utilizzo dell'aggregato, aumenta anche il rischio di esaurimento dello spazio se l'utilizzo non viene monitorato con attenzione. Inoltre, mentre si utilizza un sistema al 99% della capacità potrebbe non verificarsi un peggioramento delle performance, tuttavia si verificherebbe un sforzo di gestione che impedirebbe il riempimento completo del sistema mentre si ordina hardware aggiuntivo e potrebbe essere necessario del tempo per l'acquisto e l'installazione di dischi aggiuntivi.

Aggregati HDD, compresi gli aggregati Flash Pool

NetApp consiglia almeno il 15% di spazio libero quando si utilizzano unità rotanti. Ciò comprende tutto lo spazio inutilizzato, compreso lo spazio libero all'interno dell'aggregato o di un volume ed eventuale spazio libero allocato a causa dell'utilizzo del provisioning completo, ma non utilizzato dai dati effettivi. Le prestazioni saranno influenzate dall'avvicinarsi del 10% dello spazio libero.

Macchine virtuali di storage

La gestione dello storage del database Oracle è centralizzata su una Storage Virtual

Machine (SVM)

Una SVM, nota come vserver nell'interfaccia a riga di comando di ONTAP, è un'unità funzionale di base dello storage ed è utile confrontare una SVM con un guest su un server VMware ESX.

Quando viene installato per la prima volta, ESX non dispone di funzionalità preconfigurate, come l'hosting di un sistema operativo guest o il supporto di un'applicazione per l'utente finale. Si tratta di un container vuoto fino a quando non viene definita una macchina virtuale (VM). ONTAP è simile. Quando viene installata per la prima volta, ONTAP non dispone di funzionalità di servizio dati fino a quando non viene creata una SVM. È il linguaggio della SVM che definisce i servizi dati.

Come per altri aspetti dell'architettura dello storage, le migliori opzioni per il design di SVM e interfaccia logica (LIF) dipendono in gran parte dai requisiti di scalabilità e dalle esigenze di business.

SVM

Non esistono Best practice ufficiali per il provisioning di SVM per ONTAP. Il giusto approccio dipende dai requisiti di gestione e sicurezza.

La maggior parte dei clienti utilizza una SVM primaria per la maggior parte delle loro esigenze quotidiane, quindi crea un piccolo numero di SVM per esigenze speciali. Ad esempio, è possibile creare:

- Una SVM per un database aziendale critico gestita da un team di specialisti
- Una SVM per un gruppo di sviluppo al quale è stato assegnato un controllo amministrativo completo in modo da poter gestire il proprio storage in maniera indipendente
- Una SVM per i dati di business sensibili, come le risorse umane o i dati di reporting finanziario, per cui il team di amministrazione deve essere limitato

In un ambiente multi-tenant, è possibile assegnare a ciascun tenant una SVM dedicata. Il limite per il numero di SVM e LIF per cluster, coppia ha e nodo dipende dal protocollo in uso, dal modello di nodo e dalla versione di ONTAP. Consultare "NetApp Hardware Universe" per questi limiti.

Gestione delle performance con QoS ONTAP

La gestione sicura ed efficiente di più database Oracle richiede un'efficace strategia di QoS. Il motivo è rappresentato dalle funzionalità di performance in costante aumento offerte da un sistema storage moderno.

Nello specifico, la maggiore adozione dello storage all-flash ha permesso il consolidamento dei carichi di lavoro. Gli storage array che si affidano a supporti rotanti tendevano a supportare solo un numero limitato di workload i/o-intensive a causa delle limitate funzionalità IOPS della tecnologia delle unità rotazionali meno recente. Uno o due database altamente attivi saturerebbero i dischi sottostanti molto prima che gli storage controller raggiungano i loro limiti. Questo è cambiato. La capacità di performance di un numero relativamente contenuto di dischi SSD è in grado di saturare anche gli storage controller più potenti. Ciò significa che è possibile sfruttare tutte le funzionalità dei controller senza la paura di un improvviso crollo delle performance con picchi di latenza dei supporti rotanti.

Come esempio di riferimento, un semplice sistema ha AFF A800 a due nodi è in grado di fornire fino a un milione di IOPS casuali prima che la latenza superi un millisecondo. Ci si aspetta che pochissimi carichi di lavoro singoli raggiungano tali livelli. L'utilizzo completo di questo array di sistema AFF A800 implicherà l'hosting di più carichi di lavoro, per questo motivo in modo sicuro, garantendo al contempo la prevedibilità dei requisiti di qualità del servizio.

Esistono due tipi di qualità del servizio (QoS) in ONTAP: IOPS e larghezza di banda. È possibile applicare controlli di qualità del servizio a SVM, volumi, LUN e file.

QoS (IOPS)

Un controllo della qualità del servizio IOPS si basa ovviamente sugli IOPS totali di una data risorsa, ma esistono alcuni aspetti della qualità del servizio IOPS che potrebbero non essere intuitivi. Alcuni clienti sono rimasti colpiti dall'apparente aumento della latenza al raggiungimento di una soglia IOPS. L'aumento della latenza è il risultato naturale della limitazione degli IOPS. Logicamente, funziona in modo simile a un sistema token. Ad esempio, se un dato volume contenente file di dati ha un limite di 10K IOPS, ogni i/o che arriva deve prima ricevere un token per continuare l'elaborazione. Fino a quando non sono stati consumati più di 10K gettoni in un dato secondo, non sono presenti ritardi. Se le operazioni io devono attendere per ricevere il token, questa attesa viene visualizzata come latenza aggiuntiva. Più un carico di lavoro supera il limite di qualità del servizio, più a lungo ogni i/o deve attendere in coda per l'elaborazione del proprio turno, che appare all'utente come una latenza più elevata.

Prestare attenzione nell'applicazione dei controlli QoS ai dati dei log di transazione/ripristino del database. Mentre le richieste di performance del logging di redo sono in genere molto, molto più basse dei data afiles, l'attività del log di redo è molto bursty. L'io avviene in brevi impulsi e un limite QoS che appare appropriato per i livelli di io di redo medi potrebbe essere troppo basso per i requisiti effettivi. Il risultato può essere una serie di limitazioni delle performance, mentre la qualità del servizio viene associata a ogni burst dei log di ripristino. In generale, il redo e la registrazione dell'archivio non devono essere limitati dalla QoS.

QoS della larghezza di banda

Non tutte le dimensioni i/o sono uguali. Ad esempio, un database potrebbe eseguire un elevato numero di piccoli blocchi di lettura con il raggiungimento della soglia IOPS, tuttavia, è possibile che i database eseguano anche un'operazione di scansione completa della tabella, che consisterebbe in un numero molto ridotto di letture di blocchi di grandi dimensioni, consumando una grande quantità di larghezza di banda ma con un numero relativamente basso di IOPS.

Allo stesso modo, un ambiente VMware potrebbe gestire un numero molto elevato di IOPS casuali durante l'avvio, ma eseguirebbe un numero minore di io, ma più grande, durante un backup esterno.

Una gestione efficace delle performance a volte richiede limiti di qualità del servizio (QoS) IOPS o larghezza di banda, o anche entrambi.

Qualità del servizio minima/garantita

Molti clienti cercano una soluzione che includa QoS garantita, che sia più difficile da raggiungere di quanto possa sembrare e che sia potenzialmente abbastanza dispendiosa. Ad esempio, collocare 10 database con una garanzia di 10K IOPS richiede il dimensionamento di un sistema per uno scenario in cui tutti i 10 database vengono eseguiti contemporaneamente a 10K IOPS, per un totale di 100K.

L'utilizzo ottimale per i controlli minimi della qualità del servizio è la protezione dei carichi di lavoro critici. Ad esempio, prendi in considerazione un controller ONTAP con un numero massimo di IOPS possibile di 500K e un mix di workload di produzione e sviluppo. È consigliabile applicare policy QoS massime ai carichi di lavoro di sviluppo per impedire a qualsiasi database di monopolizzare il controller. Quindi, ai carichi di lavoro di produzione si applicano policy minime di qualità del servizio per assicurarsi che dispongano sempre degli IOPS richiesti, quando necessario.

QoS adattiva

La qualità del servizio adattiva fa riferimento alla funzionalità ONTAP, in cui il limite della qualità del servizio si basa sulla capacità dell'oggetto storage. Viene utilizzata raramente con i database perché di solito non esiste alcun collegamento tra le dimensioni di un database e i relativi requisiti prestazionali. I database di grandi dimensioni possono essere quasi inerti, mentre quelli di dimensioni inferiori possono utilizzare un numero elevato di IOPS.

La qualità del servizio adattiva può rivelarsi molto utile con i datastore di virtualizzazione, perché i requisiti di IOPS di tali set di dati tendono a correlare le dimensioni totali del database. Un datastore più recente, che contiene 1TB TB di file VMDK, avrà probabilmente bisogno di circa la metà delle performance rispetto a un datastore da 2TB TB. La qualità del servizio adattiva ti consente di aumentare automaticamente i limiti della qualità del servizio, man mano che il datastore viene popolato con i dati.

Efficienza

Le funzionalità di efficienza dello spazio di ONTAP sono ottimizzate per i database Oracle. In quasi tutti i casi, l'approccio migliore è quello di lasciare le impostazioni predefinite con tutte le funzioni di efficienza attivate.

Le funzionalità di efficienza in termini di spazio, come compressione, compaction e deduplica, sono progettate per aumentare la quantità di dati logici applicabili a una determinata quantità di storage fisico. Il risultato è una riduzione dei costi e dell'overhead di gestione.

Ad un livello elevato, la compressione è un processo matematico in cui gli schemi nei dati vengono rilevati e codificati in modo da ridurre i requisiti di spazio. La deduplica, invece, rileva i blocchi di dati effettivi e ripetuti e rimuove le copie estranee. La tecnologia di compaction consente a più blocchi logici di dati di condividere lo stesso blocco fisico sui supporti.

Per una spiegazione dell'interazione tra efficienza dello storage e prenotazione frazionata, vedere le sezioni seguenti sul thin provisioning.

Compressione

Prima della disponibilità dei sistemi storage all-flash, la compressione basata su array aveva un valore limitato, perché la maggior parte dei carichi di lavoro con i/o-intensive richiedeva un numero molto elevato di spindle per fornire performance accettabili. I sistemi storage contenevano invariabilmente una capacità superiore rispetto a quella richiesta come effetto collaterale dell'elevato numero di dischi. La situazione è cambiata con l'ascesa dello storage a stato solido. Non è più necessario effettuare un provisioning in eccesso significativo dei dischi solo per ottenere buone prestazioni. Lo spazio su disco di un sistema di storage può essere adattato alle effettive esigenze di capacità.

L'aumento della capacità degli IOPS dei dischi a stato solido (SSD) offre quasi sempre risparmi sui costi rispetto ai dischi rotanti, ma la compressione può ottenere ulteriori risparmi aumentando la capacità effettiva dei supporti a stato solido.

Esistono diversi modi per comprimere i dati. Molti database includono proprie funzionalità di compressione, sebbene raramente queste vengano osservate negli ambienti dei clienti. Il motivo è solitamente la penalizzazione delle prestazioni per una **modifica** dei dati compressi, mentre con alcune applicazioni vi sono elevati costi di licenza per la compressione a livello di database. Infine, ci sono le conseguenze globali delle performance sulle operazioni di database. Ha poco senso pagare un costo elevato di licenza per CPU per una CPU che esegue la compressione e la decompressione dei dati piuttosto che un vero lavoro di database. Un'opzione migliore è trasferire il lavoro di compressione sul sistema storage.

Compressione adattiva

La compressione adattiva è stata testata accuratamente con carichi di lavoro Enterprise senza effetti osservati sulle performance, anche in un ambiente all-flash in cui la latenza viene misurata in microsecondi. Alcuni clienti hanno anche segnalato un aumento delle performance con l'utilizzo della compressione, perché i dati rimangono compressi nella cache, aumentando di fatto la quantità di cache disponibile in un controller.

ONTAP gestisce i blocchi fisici in 4KB unità. La compressione adattiva utilizza dimensioni predefinite dei blocchi di compressione di 8KB KB, il che significa che i dati sono compressi in unità da 8KB KB. Corrisponde alle dimensioni dei blocchi di 8KB KB utilizzate più spesso dai database relazionali. Gli algoritmi di compressione diventano più efficienti con la compressione di un numero maggiore di dati come una singola unità. Una dimensione dei blocchi di compressione da 32KB KB sarebbe più efficiente in termini di spazio rispetto a un'unità dei blocchi di compressione da 8KB KB. Ciò significa che la compressione adattiva che utilizza le dimensioni predefinite dei blocchi di 8KB KB produce tassi di efficienza leggermente inferiori, ma esiste anche un vantaggio significativo nell'utilizzo di dimensioni inferiori dei blocchi di compressione. I carichi di lavoro dei database includono un'elevata attività di sovrascrittura. La sovrascrittura di un 8KB di un blocco di dati 32KB compresso richiede la lettura dell'intero 32KB di dati logici, la decompressione, l'aggiornamento della regione 8KB richiesta, la ricompressione e quindi la riscrittura dell'intero 32KB sui dischi. Si tratta di un'operazione molto costosa per un sistema storage ed è il motivo per cui alcuni storage array concorrenti basati su dimensioni dei blocchi di compressione più grandi implicano anche una significativa penalizzazione delle performance con i carichi di lavoro dei database.

Le dimensioni dei blocchi utilizzate dalla compressione adattiva possono essere aumentate fino a 32KB KB. Questo può migliorare l'efficienza di archiviazione e dovrebbe essere considerato per i file inattivi come i log delle transazioni e i file di backup quando una quantità sostanziale di tali dati è memorizzata nell'array. In alcune situazioni, i database attivi che utilizzano dimensioni blocco 16KB KB o 32KB KB possono anche trarre vantaggio dall'aumento delle dimensioni blocco della compressione adattiva per adeguarsi. Consulta un NetApp o un rappresentante del partner per ottenere indicazioni relative all'adeguatezza del tuo carico di lavoro.

Le dimensioni dei blocchi di compressione superiori a 8KB KB non devono essere utilizzate insieme alla deduplica nelle destinazioni di backup in streaming. Il motivo è che piccole modifiche ai dati di backup influiscono sulla finestra di compressione 32KB. Se la finestra si sposta, i dati compressi risultanti differiscono per l'intero file. La deduplica si verifica dopo la compressione, il che significa che il motore di deduplica vede ogni backup compresso in modo diverso. Se è richiesta la deduplica dei backup in streaming, è consigliabile utilizzare solo la compressione adattiva per blocchi da 8KB KB. La compressione adattiva è preferibile, perché funziona a blocchi di dimensioni inferiori e non interrompe l'efficienza di deduplica. Per motivi simili, la compressione lato host interferisce anche con l'efficienza della deduplica.

Allineamento delle compressioni

La compressione adattiva in un ambiente di database richiede alcune considerazioni sull'allineamento dei blocchi di compressione. Ciò rappresenta solo una preoccupazione per i dati che sono soggetti a sovrascritture casuali di blocchi molto specifici. Questo approccio è simile in teoria all'allineamento complessivo del file system, dove l'inizio di un file system deve essere allineato al limite di un dispositivo 4K e la dimensione di blocco di un file system deve essere un multiplo di 4K.

Ad esempio, una scrittura 8KB in un file viene compressa solo se si allinea con un limite 8KB all'interno del file system stesso. Questo punto significa che deve rientrare nel primo 8KB del file, nel secondo 8KB del file e così via. Il modo più semplice per garantire un corretto allineamento è utilizzare il tipo di LUN corretto, ogni partizione creata dovrebbe avere un offset dall'inizio del dispositivo che è un multiplo di 8K, e utilizzare una dimensione del blocco del file system che è un multiplo della dimensione del blocco del database.

Dati come backup o log delle transazioni sono operazioni scritte in sequenza che coprono più blocchi, tutti compressi. Pertanto, non è necessario considerare l'allineamento. L'unico modello di i/o che desta preoccupazione sono le sovrascritture casuali dei file.

Compaction dei dati

La data compaction è una tecnologia che migliora l'efficienza di compressione. Come indicato in precedenza, la sola compressione adattiva può garantire risparmi 2:1:1 al meglio, perché è limitata alla memorizzazione di un i/o da 8KB KB in un blocco WAFL da 4KB KB. I metodi di compressione con dimensioni dei blocchi maggiori garantiscono una maggiore efficienza. Tuttavia, non sono adatte per i dati che sono soggetti a piccole sovrascritture dei blocchi. La decompressione di 32KB unità di dati, l'aggiornamento di una porzione 8KB, la ricompressione e la riscrittura sui dischi crea overhead.

La data compaction opera consentendo di memorizzare più blocchi logici all'interno dei blocchi fisici. Ad esempio, un database con dati altamente comprimibili come testo o blocchi parzialmente completi può comprimere da 8KB a 1KB. Senza la compaction, quei 1KB PB di dati continuerebbero ad occupare un intero blocco da 4KB KB. Inline data compaction per memorizzare 1KB TB di dati compressi in sole 1KB:1 di spazio fisico insieme ad altri dati compressi. Non si tratta di una tecnologia di compressione, ma semplicemente di un metodo più efficiente per allocare spazio sulle unità e quindi non dovrebbe creare alcun effetto rilevabile sulle prestazioni.

Il grado di risparmio ottenuto varia. I dati già compressi o crittografati non possono in genere essere ulteriormente compressi, e pertanto tali set di dati non traggono vantaggio dalla compattazione. Al contrario, i file di dati appena inizializzati contenenti poco più dei metadati dei blocchi e la compressione di zeri fino a 80:1.

Efficienza di conservazione sensibile alla temperatura

L'efficienza di stoccaggio sensibile alla temperatura (TSSE) è disponibile in ONTAP 9.8 e versioni successive. Si affida alle mappe termiche di accesso ai blocchi per identificare i blocchi a cui si accede raramente e comprimerli con una maggiore efficienza.

Deduplica

La deduplica consiste nella rimozione di dimensioni dei blocchi duplicate da un set di dati. Ad esempio, se lo stesso blocco 4KB esistesse in 10 file diversi, la deduplica reindirizzerebbe quel blocco 4KB in tutti i file 10 allo stesso blocco fisico da 4KB KB. Il risultato sarebbe un miglioramento di 10:1 volte in efficienza per quei dati.

Dati come i LUN di avvio guest di VMware si deduplicano in genere in modo estremamente efficace poiché sono costituiti da più copie degli stessi file del sistema operativo. Sono state osservate un'efficienza pari o superiore a 100:1.

Alcuni dati non contengono dati duplicati. Ad esempio, un blocco Oracle contiene un'intestazione univoca a livello globale per il database e un trailer quasi univoco. Di conseguenza, la deduplica di un database Oracle raramente offre un risparmio superiore al 1%. La deduplica con i database MS SQL è leggermente migliore, ma i metadati univoci a livello di blocco rimangono un limite.

In pochi casi, sono stati osservati risparmi di spazio fino al 15% nei database con blocchi di dimensioni grandi e 16KB. Il 4KB iniziale di ciascun blocco contiene la testata unica a livello globale, mentre il 4KB finale contiene il rimorchio quasi unico. I blocchi interni sono candidati per la deduplica, sebbene in pratica ciò sia quasi interamente attribuito alla deduplica di dati azzerati.

Molti array della concorrenza rivendicano la capacità di deduplicare i database sulla base del presupposto che un database venga copiato più volte. Anche in questo caso è possibile utilizzare la deduplica NetApp, ma ONTAP offre un'opzione migliore: La tecnologia FlexClone di NetApp. Il risultato finale è lo stesso; vengono

create più copie di un database che condividono la maggior parte dei blocchi fisici sottostanti. L'utilizzo di FlexClone è molto più efficiente della necessità di dedicare tempo alla copia e alla deduplica dei file di database. In effetti, non viene effettuata alcuna duplicazione piuttosto che deduplica, poiché al primo posto non viene mai creato un duplicato.

Efficienza e thin provisioning

Le funzionalità di efficienza sono forme di thin provisioning. Ad esempio, una LUN da 100GB GB che occupa un volume da 100GB GB potrebbe comprimere fino a 50GB GB. Non ci sono risparmi effettivi ancora realizzati perché il volume è ancora 100GB. Le dimensioni del volume devono essere innanzitutto ridotte in modo che lo spazio salvato possa essere utilizzato in un'altra posizione del sistema. Se successivamente le modifiche apportate al LUN da 100GB GB rendono i dati meno comprimibili, il LUN aumenta le dimensioni e il volume potrebbe riempirsi.

Il thin provisioning è vivamente consigliato in quanto consente di semplificare la gestione, offrendo al contempo un sostanziale miglioramento della capacità utilizzabile con conseguenti risparmi sui costi. Il motivo è semplice: Gli ambienti di database includono spesso molto spazio vuoto, un elevato numero di volumi e LUN e dati comprimibili. Il thick provisioning crea la riserva di spazio sullo storage per volumi e LUN, nel caso in cui un giorno raggiungano il 100% di riempimento e contengano dati non comprimibili al 100%. È improbabile che ciò accada mai. Il thin provisioning consente di recuperare lo spazio e di utilizzarlo altrove e consente la gestione della capacità basata sul sistema storage stesso piuttosto che su molti volumi e LUN più piccoli.

Alcuni clienti preferiscono utilizzare il thick provisioning, per carichi di lavoro specifici o generalmente basato su pratiche operative e di approvvigionamento consolidate.

> Se un volume viene sottoposto a thick provisioning, è necessario fare attenzione a disattivare completamente tutte le funzionalità di efficienza per quel volume, inclusa la decompressione e la rimozione della deduplica tramite il `sis undo` comando. Il volume non dovrebbe comparire nell' `volume efficiency show`output. In tal caso, il volume è ancora parzialmente configurato per le funzioni di efficienza. Di conseguenza, la sovrascrittura garantisce un funzionamento diverso, aumentando le possibilità che le sovrascritture causino l'esaurimento inaspettato dello spazio del volume, con conseguenti errori di i/o del database.

Best practice di efficienza

NetApp consiglia quanto segue:

Valori predefiniti AFF

I volumi creati su ONTAP in esecuzione su un sistema AFF all-flash vengono sottoposti a thin provisioning con tutte le funzionalità di efficienza inline abilitate. Sebbene in genere i database non beneficino della deduplica e possano includere dati non comprimibili, le impostazioni predefinite sono comunque appropriate per quasi tutti i carichi di lavoro. ONTAP è progettato per elaborare in modo efficiente tutti i tipi di dati e gli schemi i/o, indipendentemente dal fatto che comportino risparmi. Le impostazioni predefinite devono essere modificate solo se le ragioni sono pienamente comprese e se vi è un vantaggio a deviare.

Raccomandazioni generali

- Se i volumi e/o le LUN non sono dotati di thin provisioning, è necessario disabilitare tutte le impostazioni di efficienza perché queste funzionalità non offrono risparmi e la combinazione del thick provisioning con l'efficienza dello spazio può causare comportamenti imprevisti, inclusi errori di spazio esaurito.
- Se i dati non sono soggetti a sovrascritture, ad esempio con i backup o i log delle transazioni dei database, puoi ottenere una maggiore efficienza abilitando TSSE con un periodo di raffreddamento ridotto.

- Alcuni file potrebbero contenere una quantità significativa di dati non comprimibili, ad esempio quando la compressione è già abilitata a livello di applicazione dei file sono crittografati. Se uno di questi scenari è vero, considerare la possibilità di disattivare la compressione per consentire un funzionamento più efficiente su altri volumi che contengono dati comprimibili.
- Non utilizzare sia la compressione 32KB che la deduplica con i backup del database. Vedere la sezione Compressione adattiva per ulteriori informazioni.

Thin provisioning

Il thin provisioning per un database Oracle richiede un'attenta pianificazione, perché ne consegue che è possibile configurare più spazio su un sistema di storage rispetto a quello necessariamente fisicamente disponibile. Vale la pena di fare tutto questo perché, se eseguito correttamente, il risultato è un notevole risparmio sui costi e un miglioramento della gestibilità.

Il thin provisioning è disponibile in molte forme e rappresenta parte integrante di molte funzionalità offerte da ONTAP a un ambiente applicativo aziendale. Il thin provisioning è inoltre strettamente correlato alle tecnologie di efficienza per lo stesso motivo: Le funzionalità di efficienza consentono di memorizzare dati più logici rispetto a quanto tecnicamente esistente nel sistema storage.

Quasi tutti gli utilizzi delle snapshot implicano il thin provisioning. Ad esempio, un tipico database da 10TB TB su storage NetApp include circa 30 giorni di snapshot. Questa disposizione risulta in circa 10TB di dati visibili nel file system attivo e 300TB dedicati agli snapshot. In genere, il 310TB GB di storage totale risiede su un totale di circa 12TB - 15TB GB di spazio. Il database attivo consuma 10TB e i restanti 300TB di dati richiedono solo da 2TB a 5TB di spazio, in quanto vengono memorizzate solo le modifiche apportate ai dati originali.

Anche il cloning è un esempio di thin provisioning. Un importante cliente NetApp ha creato 40 cloni di un database da 80TB TB per l'utilizzo da parte dello sviluppo. Se tutti i 40 sviluppatori che utilizzano questi cloni sovrascrivono ogni blocco in ogni file dati, sarebbero necessari oltre 3,2PB TB di storage. In pratica, il turnover è basso e il requisito di spazio collettivo è più vicino a 40TB, perché solo le modifiche sono memorizzate sui drive.

Gestione dello spazio

È necessario prestare particolare attenzione al thin provisioning di un ambiente applicativo, perché la velocità di modifica dei dati può aumentare inaspettatamente. Ad esempio, il consumo di spazio dovuto agli snapshot può crescere rapidamente se le tabelle di database vengono riindicizzate o se viene applicata una patch su larga scala ai guest VMware. Un backup posizionato in modo errato può scrivere una grande quantità di dati in un tempo molto breve. Infine, può essere difficile recuperare alcune applicazioni se un file system esaurisce inaspettatamente lo spazio libero.

Fortunatamente, questi rischi possono essere risolti con un'attenta configurazione di `volume-autogrow` e. `snapshot-autodelete` criteri: Come indicato dai nomi, queste opzioni consentono a un utente di creare policy in grado di liberare automaticamente lo spazio occupato dalle snapshot o di far crescere un volume per ospitare dati aggiuntivi. Sono disponibili molte opzioni e le esigenze variano in base al cliente.

Vedere "documentazione per la gestione logica dello storage" per una discussione completa di queste funzioni.

Prenotazioni frazionarie

Riserva frazionaria si riferisce al comportamento di un LUN in un volume rispetto all'efficienza dello spazio. Quando l'opzione `fractional-reserve` è impostato al 100%, tutti i dati nel volume possono subire un turnover del 100% con qualsiasi modello di dati senza esaurire lo spazio sul volume.

Ad esempio, si consideri un database su una singola LUN da 250GB GB in un volume da 1TB GB. La creazione di uno snapshot comporterebbe immediatamente la riserva di ulteriori 250GB GB di spazio nel volume per garantire che il volume non esaurisca lo spazio per alcun motivo. L'utilizzo di riserve frazionarie comporta in genere uno spreco di risorse poiché è estremamente improbabile che ogni byte nel volume del database debba essere sovrascritto. Non c'è motivo di riservare spazio per un evento che non si verifica mai. Tuttavia, se un cliente non è in grado di monitorare il consumo di spazio in un sistema di storage e deve essere certo che lo spazio non si esaurisce mai, sarebbero necessarie prenotazioni frazionarie del 100% per utilizzare gli snapshot.

Compressione e deduplica

Compressione e deduplica sono entrambe forme di thin provisioning. Ad esempio, un impatto dei dati di 50TB:1 potrebbe comprimere fino a 30TB:1, ottenendo un risparmio di 20TB:1. Affinché la compressione possa produrre vantaggi, alcuni di questi 20TB TB devono essere utilizzati per altri dati, altrimenti il sistema storage deve essere acquistato con meno di 50TB TB. In questo modo è possibile memorizzare una quantità di dati superiore rispetto a quella tecnicamente disponibile sul sistema storage. Dal punto di vista dei dati, i dati sono 50TB, anche se occupano solo 30TB sulle unità.

Esiste sempre la possibilità che la compressibilità di un set di dati cambi, con conseguente aumento del consumo di spazio reale. Questo aumento dei consumi implica che la compressione deve essere gestita come con altre forme di thin provisioning in termini di monitoraggio e utilizzo `volume-autogrow` e. `snapshot-autodelete`.

Compressione e deduplica sono descritte in dettaglio nella sezione xref:./oracle/efficiency.html

Compressioni e prenotazioni frazionarie

La compressione è una forma di thin provisioning. Le prenotazioni frazionarie influiscono sull'utilizzo della compressione, con una nota importante; lo spazio viene riservato prima della creazione dell'istantanea. Normalmente, la riserva frazionaria è importante solo se esiste uno snapshot. Se non è presente uno snapshot, la riserva frazionaria non è importante. Questo non è il caso della compressione. Se viene creata una LUN su un volume con compressione, ONTAP preserva lo spazio per ospitare uno snapshot. Questo comportamento può creare confusione durante la configurazione, ma è previsto.

Ad esempio, consideriamo un volume da 10GB GB con una LUN da 5GB GB compressa a 2,5GB GB senza snapshot. Considerare questi due scenari:

- Riserva frazionaria = 100 risultati in 7,5GB utilizzo
- Riserva frazionaria = 0 risultati in 2,5GB utilizzo

Il primo scenario include 2,5GB di consumo di spazio per i dati attuali e 5GB di spazio per rappresentare il 100% di fatturato della fonte in previsione dell'utilizzo di snapshot. Il secondo scenario non riserva spazio aggiuntivo.

Sebbene questa situazione possa sembrare confusa, è improbabile che si verifichi nella pratica. La compressione implica thin provisioning e il thin provisioning in un ambiente LUN richiede prenotazioni frazionarie. È sempre possibile sovrascrivere i dati compressi con qualcosa di non comprimibile, il che significa che un volume deve essere sottoposto a thin provisioning per la compressione per consentire qualsiasi risparmio.

NetApp consiglia le seguenti configurazioni riservate:

- Impostare `fractional-reserve` a 0 quando è in atto il monitoraggio della capacità di base con `volume-autogrow` e. `snapshot-autodelete`.
- Impostare `fractional-reserve` a 100 se non vi è alcuna capacità di monitoraggio o se è impossibile scaricare lo spazio in qualsiasi circostanza.

Spazio libero e allocazione di spazio LVM

L'efficienza del thin provisioning delle LUN attive in un ambiente di file system può andare persa nel tempo man mano che i dati vengono eliminati. A meno che i dati eliminati non vengano sovrascritti con degli zeri (vedere anche ASMRU) o che lo spazio non venga liberato con il recupero dello spazio TRIM/UNMAP, i dati "eliminati" occupano sempre più spazio vuoto non allocato nel file system. Inoltre, il provisioning sottile delle LUN attive è di utilità limitata in molti ambienti di database perché i file di dati vengono inizializzati alla loro dimensione completa al momento della creazione.

Un'attenta pianificazione della configurazione LVM può migliorare l'efficienza e ridurre al minimo la necessità di provisioning dello storage e di ridimensionamento delle LUN. Quando si utilizza un LVM come Veritas VxVM o Oracle ASM, le LUN sottostanti vengono suddivise in estensioni che vengono utilizzate solo quando necessario. Ad esempio, se un set di dati inizia a 2TB TB ma potrebbe crescere fino a 10TB TB con il passare del tempo, è possibile inserire il set di dati in 10TB LUN con thin provisioning organizzati in un gruppo di dischi LVM. Occupa solo 2TB GB di spazio al momento della creazione e richiederebbe spazio aggiuntivo solo se le estensioni sono allocate per ospitare la crescita dei dati. Questo processo è sicuro finché lo spazio è monitorato.

Failover/switchover di ONTAP

Le operazioni di takeover e switchover dello storage devono garantire l'integrità delle operazioni dei database Oracle. Inoltre, gli argomenti utilizzati dalle operazioni di takeover e switchover possono influire sull'integrità dei dati se utilizzati in modo errato.

- In condizioni normali, le scritture in arrivo su un dato controller vengono mirrorate in modo sincrono per il partner. In un ambiente NetApp MetroCluster, le scritture vengono anche mirrorate su un controller remoto. Fino a quando non viene memorizzata in un supporto non volatile in tutte le posizioni, la scrittura non viene riconosciuta all'applicazione host.
- Il supporto che memorizza i dati di scrittura è chiamato memoria non volatile o NVMEM. Viene anche talvolta indicata come memoria non volatile ad accesso casuale (NVRAM, nonvolatile Random Access Memory), e può essere considerata come una cache di scrittura anche se funziona come un journal. In condizioni normali, i dati provenienti da NVMEM non vengono letti; vengono utilizzati solo per proteggere i dati in caso di guasti software o hardware. Quando i dati vengono scritti sulle unità disco rigido, i dati vengono trasferiti dalla RAM nel sistema, non da NVMEM.
- Durante un'operazione di takeover, un nodo di una coppia ha (high Availability) assume il controllo delle operazioni dal partner. Lo switchover è praticamente identico, ma si applica alle configurazioni MetroCluster in cui un nodo remoto assume le funzioni di un nodo locale.

Durante le normali operazioni di manutenzione, un'operazione di takeover o switchover dello storage deve essere trasparente, ad eccezione di una potenziale breve pausa nelle operazioni in base al cambiamento dei percorsi di rete. Il networking può rivelarsi complesso, tuttavia, ed è facile commettere errori, pertanto NetApp consiglia vivamente di eseguire accuratamente le operazioni di takeover e switchover prima di mettere in produzione un sistema storage. In questo modo, è possibile verificare che tutti i percorsi di rete siano configurati correttamente. In un ambiente SAN, controllare attentamente l'output del comando `sanlun lun`

`show -p` per assicurarsi che tutti i percorsi primario e secondario previsti siano disponibili.

Occorre prestare attenzione quando si rilascia un'acquisizione forzata o uno switchover. Imporre una modifica alla configurazione dello storage con queste opzioni significa che lo stato del controller proprietario delle unità non viene preso in considerazione e il nodo alternativo assume forzatamente il controllo delle unità. Una forzatura non corretta di un takeover può causare la perdita o il danneggiamento dei dati. Questo perché un takeover o uno switchover forzato può scartare il contenuto di NVMEM. Una volta completato il takeover o lo switchover, la perdita dei dati potrebbe riportare i dati memorizzati nelle unità a uno stato leggermente più vecchio dal punto di vista del database.

Raramente dovrebbe essere necessario un takeover forzato con una normale coppia ha. In quasi tutti gli scenari di errore, un nodo si arresta e informa il partner in modo che si verifichi un failover automatico. In alcuni casi, ad esempio in caso di guasto permanente che causa la perdita dell'interconnessione tra i nodi e la perdita di un controller, è necessario eseguire un takeover forzato. In una situazione del genere, il mirroring tra i nodi viene perso prima del guasto del controller, il che significa che il controller rimasto non avrebbe più una copia delle scritture in corso. Il takeover deve quindi essere forzato, il che significa che potenzialmente i dati vengono persi.

La stessa logica si applica a uno switchover MetroCluster. In condizioni normali, lo switchover è quasi trasparente. Tuttavia, un disastro può causare una perdita di connettività tra il sito rimasto e il sito disastroso. Dal punto di vista del sito sopravvissuto, il problema potrebbe essere nient'altro che un'interruzione della connettività tra i siti, e il sito originale potrebbe ancora elaborare i dati. Se un nodo non è in grado di verificare lo stato del controller primario, è possibile solo uno switchover forzato.

NetApp raccomanda adottare le seguenti precauzioni:

- Prestare molta attenzione a non forzare accidentalmente un'acquisizione o uno switchover. In genere, non è necessario forzare e forzare la modifica può causare la perdita di dati.
- Se è necessario un takeover o uno switchover forzato, assicurarsi che le applicazioni vengano arrestate, che tutti i file system vengano dismontati e che i gruppi di volumi LVM (Logical Volume Manager) siano diversi. I gruppi di dischi ASM devono essere smontati.
- In caso di switchover MetroCluster forzato, scollegare il nodo guasto da tutte le risorse di storage rimaste. Per ulteriori informazioni, consultare la Guida alla gestione e al ripristino di emergenza di MetroCluster per la versione pertinente di ONTAP.

MetroCluster e aggregati multipli

MetroCluster è una tecnologia di replica sincrona che passa alla modalità asincrona in caso di interruzione della connettività. Questa è la richiesta più comune da parte dei clienti, perché la replica sincrona garantita significa che l'interruzione della connettività del sito porta a uno stallo completo dell'i/o del database, mettendo il database fuori servizio.

Con MetroCluster, gli aggregati vengono sincronizzati rapidamente dopo il ripristino della connettività. A differenza di altre tecnologie di storage, MetroCluster non dovrebbe mai richiedere un reindirizzamento completo in seguito a un guasto del sito. È necessario spedire solo le modifiche delta.

Nei set di dati estesi agli aggregati c'è solo un piccolo rischio che occorrano passaggi aggiuntivi di recovery dei dati in uno scenario di emergenza regolare. In particolare, se (a) la connettività tra siti viene interrotta, (b) la connettività viene ripristinata, (c) gli aggregati raggiungono uno stato in cui alcuni sono sincronizzati e alcuni non lo sono, quindi (d) il sito primario viene perso, il risultato è un sito sopravvissuto in cui gli aggregati non sono sincronizzati tra loro. In tal caso, parti del set di dati vengono sincronizzate tra loro e non è possibile ripristinare applicazioni, database o datastore senza un recovery. Se un set di dati si estende agli aggregati, NetApp consiglia vivamente di sfruttare i backup basati su snapshot con uno dei molti strumenti disponibili, per

verificare la possibilità di recupero rapido in questo scenario insolito.

Configurazione ONTAP sui sistemi ASA r2

RAID

RAID si riferisce all'uso della ridondanza basata sulla parità per proteggere i dati dai guasti delle unità. ASA r2 utilizza le stesse tecnologie ONTAP RAID dei sistemi AFF e FAS , garantendo una protezione solida contro i guasti di più dischi.

ONTAP esegue automaticamente la configurazione RAID per i sistemi ASA r2. Si tratta di un componente fondamentale dell'esperienza di gestione dello storage semplificata introdotta con la personalità ASA r2.

I dettagli chiave riguardanti la configurazione RAID automatica su ASA r2 includono:

- Zone di disponibilità dello storage (SAZ): anziché gestire manualmente gli aggregati tradizionali e i gruppi RAID, ASA r2 utilizza le zone di disponibilità dello storage (SAZ). Si tratta di pool di dischi condivisi e protetti da RAID per una coppia HA, in cui entrambi i nodi hanno accesso completo allo stesso storage.
- Posizionamento automatico: quando viene creata un'unità di archiviazione (LUN o spazio dei nomi NVMe), ONTAP crea automaticamente un volume all'interno della SAZ e lo posiziona per ottenere prestazioni ottimali e un equilibrio di capacità.
- Nessuna gestione aggregata manuale: i comandi tradizionali di gestione degli aggregati e dei gruppi RAID non sono supportati su ASA r2. In questo modo gli amministratori non devono più pianificare manualmente le dimensioni dei gruppi RAID, i dischi di parità o le assegnazioni dei nodi.
- Provisioning semplificato: il provisioning viene gestito tramite System Manager o comandi CLI semplificati che si concentrano sulle unità di archiviazione anziché sul layout RAID fisico sottostante.
- Ribilanciamento del carico di lavoro: a partire dalle versioni 2025 (ONTAP 9.17.1), ONTAP ribilancia automaticamente i carichi di lavoro tra i nodi nella coppia HA per garantire che le prestazioni e l'utilizzo dello spazio rimangano bilanciati senza intervento manuale.

ASA r2 utilizza automaticamente le tecnologie RAID predefinite di ONTAP: RAID DP per la maggior parte delle configurazioni e RAID-TEC per pool SSD molto grandi. In questo modo si elimina la necessità di selezionare manualmente il RAID. Questi livelli RAID basati sulla parità garantiscono una maggiore efficienza e affidabilità di archiviazione rispetto al mirroring, spesso consigliato dalle vecchie best practice di Oracle ma non rilevante per ASA r2. ONTAP evita la tradizionale penalità di scrittura RAID tramite l'integrazione WAFL , garantendo prestazioni ottimali per carichi di lavoro Oracle quali redo logging e scritture casuali di file di dati. In combinazione con la gestione RAID automatizzata e le Storage Availability Zone, ASA r2 garantisce elevata disponibilità e protezione di livello aziendale per i database Oracle.

Gestione della capacità

La gestione di un database o di un'altra applicazione aziendale con storage aziendale prevedibile, gestibile e ad alte prestazioni richiede spazio libero sulle unità per la gestione di dati e metadati. La quantità di spazio libero richiesta dipende dal tipo di unità utilizzata e dai processi aziendali.

ASA r2 utilizza le Storage Availability Zone (SAZ) anziché gli aggregati, ma il principio rimane lo stesso: lo spazio libero include qualsiasi capacità fisica non consumata da dati effettivi, snapshot o overhead di sistema. Bisogna considerare anche il thin provisioning: le allocazioni logiche non riflettono il reale utilizzo fisico.

Di seguito sono riportate le raccomandazioni NetApp per i sistemi di storage ASA r2 utilizzati per le applicazioni aziendali:

Pool SSD nei sistemi ASA r2

* NetApp consiglia* di mantenere almeno il 10% di spazio fisico libero negli ambienti ASA r2. Questa linea guida si applica ai pool solo SSD utilizzati dai sistemi ASA r2 e include tutto lo spazio inutilizzato all'interno delle unità SAZ e di archiviazione. Lo spazio logico non è importante; l'attenzione è rivolta allo spazio fisico effettivamente libero disponibile per l'archiviazione dei dati.

Sebbene ASA r2 possa sostenere un utilizzo elevato senza degrado delle prestazioni, il funzionamento a piena capacità aumenta il rischio di esaurimento dello spazio e di sovraccarico amministrativo durante l'espansione dello storage. Un utilizzo superiore al 90% potrebbe non avere alcun impatto sulle prestazioni, ma potrebbe complicare la gestione e ritardare il provisioning di unità aggiuntive.

I sistemi ASA r2 supportano unità di storage fino a 128 TB e dimensioni SAZ fino a 2 PB per coppia HA, con ONTAP che bilancia automaticamente la capacità tra i nodi. Il monitoraggio dell'utilizzo a livello di cluster, SAZ e unità di storage è essenziale per garantire spazio libero adeguato per snapshot, carichi di lavoro con thin provisioning e crescita futura. Se la capacità si avvicina a soglie critiche (~ 90% di utilizzo), è necessario aggiungere ulteriori SSD in gruppi (minimo sei unità) per mantenere prestazioni e resilienza.

Macchine virtuali di storage

Anche la gestione dell'archiviazione del database Oracle sui sistemi ASA r2 è centralizzata su una Storage Virtual Machine (SVM), nota come vserver nella CLI ONTAP .

Una SVM è l'unità fondamentale per il provisioning e la sicurezza dello storage in ONTAP, simile a una VM guest su un server VMware ESX. Quando ONTAP viene installato per la prima volta su ASA r2, non ha alcuna capacità di elaborazione dati finché non viene creata una SVM. L'SVM definisce la personalità e i servizi dati per l'ambiente SAN.

I sistemi ASA r2 utilizzano una personalità ONTAP solo SAN, ottimizzata per supportare protocolli a blocchi (FC, iSCSI, NVMe/FC, NVMe/TCP) e rimuove le funzionalità correlate a NAS. Ciò semplifica la gestione e garantisce che tutte le configurazioni SVM siano ottimizzate per i carichi di lavoro SAN. A differenza dei sistemi AFF/ FAS , ASA r2 non espone opzioni per servizi NAS quali directory home o condivisioni NFS.

Quando viene creato un cluster, ASA r2 fornisce automaticamente una SVM dati predefinita denominata svm1 con protocolli SAN abilitati. Questa SVM è pronta per operazioni di archiviazione a blocchi senza richiedere la configurazione manuale dei servizi di protocollo. Per impostazione predefinita, i LIF dei dati IP in questa SVM supportano i protocolli iSCSI e NVMe/TCP e utilizzano la policy di servizio default-data-blocks, che semplifica la configurazione iniziale per i carichi di lavoro SAN. In seguito gli amministratori possono creare SVM aggiuntive o personalizzare le configurazioni LIF in base ai requisiti di prestazioni, sicurezza o multi-tenant.

Le interfacce logiche (LIF) per i protocolli SAN devono essere progettate in base ai requisiti di prestazioni e disponibilità. ASA r2 supporta iSCSI, FC e NVMe LIF, ma tieni presente che il failover automatico iSCSI LIF non è abilitato per impostazione predefinita perché ASA r2 utilizza una rete condivisa per gli host NVMe e SCSI. Per abilitare il failover automatico, creare "LIF solo iSCSI".

SVM

Come per altre piattaforme ONTAP , non esiste una best practice ufficiale per il numero di SVM da creare; la decisione dipende dai requisiti di gestione e sicurezza.

La maggior parte dei clienti utilizza un singolo SVM primario per le operazioni quotidiane e crea SVM aggiuntivi per esigenze speciali, come:

- Un SVM dedicato per un database aziendale critico gestito da un team di specialisti
- Un SVM per un gruppo di sviluppo con controllo amministrativo delegato
- Un SVM per dati sensibili che richiedono un accesso amministrativo limitato

Negli ambienti multi-tenant, a ciascun tenant può essere assegnata una SVM dedicata. Il limite per il numero di SVM e LIF per cluster, coppia HA e nodo dipende dal protocollo utilizzato, dal modello di nodo e dalla versione di ONTAP. Consultare il "NetApp Hardware Universe" per questi limiti.

ASA r2 supporta fino a 256 SVM per cluster e per coppia HA a partire da ONTAP 9.18.1 (in precedenza 32 nelle versioni precedenti).

Gestione delle prestazioni con ONTAP QoS sui sistemi ASA r2

Per gestire in modo sicuro ed efficiente più database Oracle su ASA r2 è necessaria una strategia QoS efficace. Ciò è particolarmente importante perché i sistemi ASA r2 sono piattaforme SAN all-flash progettate per prestazioni estremamente elevate e consolidamento dei carichi di lavoro.

Un numero relativamente piccolo di SSD può saturare anche i controller più potenti, pertanto i controlli QoS sono essenziali per garantire prestazioni prevedibili su più carichi di lavoro. Come riferimento, i sistemi ASA r2 come ASA A1K o A90 possono fornire da centinaia di migliaia a oltre un milione di IOPS con una latenza inferiore al millisecondo. Sono pochissimi i carichi di lavoro singoli che consumano questo livello di prestazioni, quindi il pieno utilizzo in genere comporta l'hosting di più database o applicazioni. Per farlo in modo sicuro sono necessarie policy QoS per evitare conflitti di risorse.

ONTAP QoS su ASA r2 funziona allo stesso modo dei sistemi AFF/ FAS , con due tipi principali di controlli: IOPS e larghezza di banda. I controlli QoS possono essere applicati a SVM e LUN.

QoS (IOPS)

La QoS basata su IOPS limita gli IOPS totali per una determinata risorsa. In ASA r2, i criteri QoS possono essere applicati a livello SVM e a singoli oggetti di archiviazione, come le LUN. Quando un carico di lavoro raggiunge il limite di IOPS, vengono messe in coda ulteriori richieste di I/O per i token, il che introduce latenza. Si tratta di un comportamento previsto che impedisce a un singolo carico di lavoro di monopolizzare le risorse del sistema.

Prestare attenzione quando si applicano controlli QoS ai dati del registro delle transazioni/redo del database. Questi carichi di lavoro sono soggetti a raffiche e un limite QoS che sembra ragionevole per un'attività media potrebbe essere troppo basso per i picchi, causando gravi problemi di prestazioni. In generale, la registrazione delle operazioni di redo e di archiviazione non dovrebbe essere limitata dalla qualità del servizio.

QoS della larghezza di banda

La QoS basata sulla larghezza di banda limita la velocità di trasmissione in Mbps. Questa funzionalità è utile quando i carichi di lavoro eseguono letture o scritture di blocchi di grandi dimensioni, come scansioni complete di tabelle o operazioni di backup, che consumano una larghezza di banda significativa ma relativamente pochi IOPS. Combinando i limiti di IOPS e larghezza di banda è possibile ottenere un controllo più granulare.

Qualità del servizio minima/garantita

Le policy QoS minime riservano le prestazioni ai carichi di lavoro critici. Ad esempio, in un ambiente misto con database di produzione e sviluppo, applicare la massima qualità del servizio ai carichi di lavoro di sviluppo e la minima qualità del servizio ai carichi di lavoro di produzione per garantire prestazioni prevedibili.

QoS adattiva

La QoS adattiva regola i limiti in base alle dimensioni dell'oggetto di archiviazione. Sebbene raramente utilizzato per i database (perché le dimensioni non sono correlate alle esigenze di prestazioni), può essere utile per i carichi di lavoro di virtualizzazione in cui i requisiti di prestazioni aumentano con la capacità.

Efficienza

Le funzionalità di efficienza dello spazio ONTAP sono completamente supportate e ottimizzate per i sistemi ASA r2. Nella maggior parte dei casi, l'approccio migliore è quello di lasciare le impostazioni predefinite, con tutte le funzionalità di efficienza abilitate.

I sistemi ASA r2 sono piattaforme SAN all-flash, pertanto tecnologie di efficienza quali compressione, compattazione e deduplicazione sono fondamentali per massimizzare la capacità utilizzabile e ridurre i costi.

Compressione

La compressione riduce i requisiti di spazio codificando i modelli nei dati. Con i sistemi ASA r2 basati su SSD, la compressione garantisce risparmi significativi perché la tecnologia flash elimina la necessità di un overprovisioning per le prestazioni. La compressione adattiva ONTAP è abilitata per impostazione predefinita ed è stata ampiamente testata con carichi di lavoro aziendali, inclusi i database Oracle, senza alcun impatto misurabile sulle prestazioni, anche in ambienti in cui la latenza viene misurata in microsecondi. In alcuni casi, le prestazioni migliorano perché i dati compressi occupano meno spazio nella cache.

> L'efficienza di stoccaggio sensibile alla temperatura (TSSE) non viene applicata ai sistemi ASA r2. Nei sistemi ASA r2, la compressione non si basa su dati caldi (a cui si accede frequentemente) o freddi (a cui si accede raramente). La compressione inizia senza attendere che i dati diventino freddi.

Compressione adattiva

Per impostazione predefinita, la compressione adattiva utilizza una dimensione di blocco di 8 KB, corrispondente alla dimensione di blocco comunemente utilizzata dai database relazionali. Blocchi di dimensioni maggiori (16 KB o 32 KB) possono migliorare l'efficienza per dati sequenziali quali registri delle transazioni o backup, ma devono essere utilizzati con cautela per i database attivi per evitare sovraccarichi durante le sovrascritture.

> La dimensione del blocco può essere aumentata fino a 32 KB per i file quiescenti come registri o backup. Prima di modificare le impostazioni predefinite, consultare le istruzioni NetApp .

Non utilizzare la compressione a 32 KB con deduplicazione per i backup in streaming. Utilizzare la compressione a 8 KB per mantenere l'efficienza della deduplicazione.

Allineamento delle compressioni

L'allineamento della compressione è importante per le sovrascritture casuali. Assicurarsi che il tipo di LUN, l'offset della partizione (multiplo di 8 KB) e la dimensione del blocco del file system siano corretti e allineati alla dimensione del blocco del database. I dati sequenziali, come backup o registri, non richiedono considerazioni di allineamento.

Compaction dei dati

La compattazione integra la compressione consentendo a più blocchi compressi di condividere lo stesso blocco fisico. Ad esempio, se un blocco da 8 KB viene compresso a 1 KB, la compattazione garantisce che lo spazio rimanente non venga sprecato. Questa funzionalità è in linea e non comporta penalizzazioni in termini di prestazioni.

Deduplica

La deduplicazione rimuove i blocchi duplicati nei set di dati. Mentre i database Oracle in genere generano risparmi minimi sulla deduplicazione grazie a intestazioni e trailer di blocchi univoci, la deduplicazione ONTAP può comunque recuperare spazio da blocchi azzerati e modelli ripetuti.

Efficienza e thin provisioning

I sistemi ASA r2 utilizzano il thin provisioning per impostazione predefinita. Le funzionalità di efficienza completano il thin provisioning per massimizzare la capacità utilizzabile.

Le unità di storage sono sempre scarsamente provisionate sui sistemi di storage ASA r2. Il provisioning spesso non è supportato.

Tecnologia QuickAssist (QAT)

Nelle piattaforme NetApp ASA r2, la tecnologia Intel QuickAssist (QAT) offre un'efficienza accelerata dall'hardware che differisce in modo significativo dalla tecnologia TSSE (Temperature-Sensitive Storage Efficiency) basata su software e senza QAT.

QAT con accelerazione hardware:

- Scarica le attività di compressione e crittografia dai core della CPU.
- Consente un'efficienza immediata e in linea sia per i dati attivi (a cui si accede frequentemente) sia per quelli inattivi (a cui si accede raramente).
- Riduce significativamente il sovraccarico della CPU.
- Offre una maggiore produttività e una minore latenza.
- Migliora la scalabilità per operazioni che richiedono prestazioni elevate, come la crittografia TLS e VPN.

TSSE senza QAT:

- Si basa su processi basati sulla CPU per operazioni efficienti.
- Applica l'efficienza solo ai dati freddi dopo un ritardo.

- Consuma più risorse della CPU.
- Limita le prestazioni complessive rispetto ai sistemi accelerati da QAT.

I moderni sistemi ASA r2 garantiscono quindi un'efficienza più rapida, accelerata dall'hardware e un migliore utilizzo del sistema rispetto alle vecchie piattaforme basate solo su TSSE.

Le migliori pratiche di efficienza per ASA r2

NetApp consiglia quanto segue:

Valori predefiniti ASA r2

Le unità di storage create su ONTAP in esecuzione su sistemi ASA r2 sono sottoposte a thin provisioning con tutte le funzionalità di efficienza in linea abilitate per impostazione predefinita, tra cui compressione, compattazione e deduplicazione. Sebbene i database Oracle in genere non traggano grandi vantaggi dalla deduplicazione e possano includere dati non comprimibili, queste impostazioni predefinite sono adatte a quasi tutti i carichi di lavoro. ONTAP è progettato per elaborare in modo efficiente tutti i tipi di dati e modelli di I/O, indipendentemente dal fatto che comportino o meno risparmi. Le impostazioni predefinite dovrebbero essere modificate solo se le ragioni sono pienamente comprese e se vi è un chiaro vantaggio nel discostarsi da esse.

Raccomandazioni generali

- Disattiva la compressione per i dati crittografati o compressi dall'app: se i file sono già compressi a livello di applicazione o crittografati, disattiva la compressione per ottimizzare le prestazioni e consentire un funzionamento più efficiente su altre unità di archiviazione.
- Evitare di combinare grandi blocchi di compressione con la deduplicazione: non utilizzare sia la compressione a 32 KB che la deduplicazione per i backup del database. Per i backup in streaming, utilizzare la compressione a 8 KB per mantenere l'efficienza della deduplicazione.
- Monitoraggio dei risparmi in termini di efficienza: utilizzare gli strumenti ONTAP (System Manager, Active IQ) per monitorare gli effettivi risparmi di spazio e, se necessario, adattare le politiche.

Thin provisioning

Il thin provisioning per un database Oracle su ASA r2 richiede un'attenta pianificazione perché implica la configurazione di uno spazio logico maggiore di quello fisicamente disponibile. Se implementato correttamente, il thin provisioning garantisce notevoli risparmi sui costi e una migliore gestibilità.

Il thin provisioning è parte integrante di ASA r2 ed è strettamente correlato alle tecnologie di efficienza ONTAP , poiché entrambe consentono di archiviare più dati logici rispetto alla capacità fisica del sistema. I sistemi ASA r2 sono solo SAN e il thin provisioning si applica alle unità di archiviazione e alle LUN all'interno delle Storage Availability Zone (SAZ).

> Per impostazione predefinita, le unità di archiviazione ASA r2 sono sottoposte a thin provisioning.

Quasi tutti gli utilizzi degli snapshot prevedono il thin provisioning. Ad esempio, un tipico database da 10 TiB con 30 giorni di snapshot potrebbe apparire come 310 TiB di dati logici, ma vengono consumati solo 12-15 TiB di spazio fisico, perché gli snapshot memorizzano solo i blocchi modificati.

Allo stesso modo, la clonazione è un'altra forma di thin provisioning. Un ambiente di sviluppo con 40 cloni di un database da 80 TiB richiederebbe 3,2 PiB se fosse scritto completamente, ma in pratica ne consuma molto

meno perché vengono memorizzate solo le modifiche.

Gestione dello spazio

È necessario prestare particolare attenzione al thin provisioning in un ambiente applicativo, poiché la velocità di modifica dei dati può aumentare in modo imprevisto. Ad esempio, il consumo di spazio dovuto agli snapshot può aumentare rapidamente se le tabelle del database vengono reindicizzate o se vengono applicate patch su larga scala ai guest VMware. Un backup fuori posto può causare la perdita di una grande quantità di dati in pochissimo tempo. Infine, può essere difficile ripristinare alcune applicazioni se una LUN esaurisce inaspettatamente lo spazio libero.

In ASA r2, questi rischi vengono mitigati tramite **thin provisioning**, **monitoraggio proattivo** e **politiche di ridimensionamento LUN**, anziché tramite funzionalità ONTAP come volume-autogrow o snapshot-autodelete. Gli amministratori dovrebbero:

- Abilitare il thin provisioning sulle LUN (`space-reserve disabled`) - questa è l'impostazione predefinita in ASA r2
- Monitorare la capacità utilizzando gli avvisi di System Manager o l'automazione basata su API
- Utilizzare il ridimensionamento LUN pianificato o programmato per adattarsi alla crescita
- Configurare la riserva di snapshot e l'eliminazione automatica degli snapshot tramite System Manager (GUI)

> È essenziale pianificare attentamente le soglie di spazio e gli script di automazione perché ASA r2 non supporta la crescita automatica del volume o l'eliminazione degli snapshot tramite CLI.

ASA r2 non utilizza impostazioni di riserva frazionaria perché è un'architettura solo SAN che astrae le opzioni di volume basate su WAFL. Invece, l'efficienza dello spazio e la protezione da sovrascrittura vengono gestite a livello LUN. Ad esempio, se si dispone di una LUN da 250 GiB fornita da un'unità di archiviazione, gli snapshot consumano spazio in base alle effettive modifiche dei blocchi anziché riservare in anticipo una quantità di spazio equivalente. In questo modo si elimina la necessità di grandi prenotazioni statiche, comuni negli ambienti ONTAP tradizionali che utilizzavano la riserva frazionaria.

> Se è richiesta una protezione da sovrascrittura garantita e il monitoraggio non è fattibile, gli amministratori devono predisporre una capacità sufficiente nell'unità di archiviazione e impostare opportunamente la riserva di snapshot. Tuttavia, la progettazione di ASA r2 rende la riserva frazionaria non necessaria per la maggior parte dei carichi di lavoro.

Compressione e deduplica

La compressione e la deduplicazione in ASA r2 sono tecnologie di efficienza dello spazio, non meccanismi tradizionali di thin provisioning. Queste funzionalità riducono l'ingombro fisico dello storage eliminando i dati ridondanti e comprimendo i blocchi, consentendo di archiviare più dati logici di quanto la capacità grezza consentirebbe altrimenti.

Ad esempio, un set di dati da 50 TiB potrebbe essere compresso a 30 TiB, risparmiando 20 TiB di spazio fisico. Dal punto di vista applicativo, ci sono ancora 50 TiB di dati, anche se occupano solo 30 TiB sul disco.

> La comprimibilità di un set di dati può cambiare nel tempo, il che può aumentare il consumo di spazio fisico. Pertanto, la compressione e la deduplicazione devono essere gestite in modo proattivo attraverso il monitoraggio e la pianificazione della capacità.

Spazio libero e allocazione di spazio LVM

Il thin provisioning negli ambienti ASA r2 può perdere efficienza nel tempo se i blocchi eliminati non vengono recuperati. A meno che lo spazio non venga liberato tramite TRIM/UNMAP o sovrascritto con zeri (tramite ASMRU - Automatic Space Management and Reclamation Utility), i dati eliminati continuano a consumare capacità fisica. In molti ambienti di database Oracle, il thin provisioning offre vantaggi limitati perché i file di dati vengono solitamente pre-allocati alla loro dimensione completa durante la creazione.

Un'attenta pianificazione della configurazione LVM può migliorare l'efficienza e ridurre al minimo la necessità di provisioning dello storage e ridimensionamento delle LUN. Quando si utilizza un LVM come Veritas VxVM o Oracle ASM, le LUN sottostanti vengono suddivise in estensioni che vengono utilizzate solo quando necessario. Ad esempio, se un set di dati inizia con una dimensione di 2 TiB ma può crescere fino a 10 TiB nel tempo, questo set di dati potrebbe essere posizionato su 10 TiB di LUN con thin provisioning organizzate in un diskgroup LVM. Occuperebbe solo 2 TiB di spazio al momento della creazione e richiederebbe spazio aggiuntivo solo man mano che le estensioni vengono allocate per far fronte alla crescita dei dati. Questo processo è sicuro finché lo spazio è monitorato.

Failover ONTAP

È necessaria la conoscenza delle funzioni di acquisizione dello storage per garantire che le operazioni del database Oracle non vengano interrotte durante queste operazioni. Inoltre, gli argomenti utilizzati dalle operazioni di acquisizione possono compromettere l'integrità dei dati se utilizzati in modo errato.

In condizioni normali, le scritture in arrivo su un determinato controller vengono replicate in modo sincrono sul suo partner HA. In un ambiente ASA r2 con SnapMirror Active Sync (SM-as), le scritture vengono anche replicate su un controller remoto nel sito secondario. Finché una scrittura non viene memorizzata su supporti non volatili in tutte le posizioni, non viene riconosciuta dall'applicazione host.

Il supporto su cui vengono memorizzati i dati di scrittura è denominato memoria non volatile (NVMEM). A volte viene definita memoria non volatile ad accesso casuale (NVRAM) e può essere considerata più un registro di scrittura che una cache. Durante il normale funzionamento, i dati provenienti NVMEM non vengono letti; vengono utilizzati solo per proteggere i dati in caso di guasto del software o dell'hardware. Quando i dati vengono scritti sulle unità, vengono trasferiti dalla RAM di sistema e non dalla NVMEM.

Durante un'operazione di acquisizione, un nodo in una coppia HA assume il controllo delle operazioni dal partner. In ASA r2, lo switchover non è applicabile perché MetroCluster non è supportato; al suo posto, SnapMirror Active Sync fornisce ridondanza a livello di sito. Le operazioni di acquisizione dello storage durante la manutenzione ordinaria dovrebbero essere trasparenti, fatta eccezione per una breve pausa nelle operazioni quando cambiano i percorsi di rete. Il networking può essere complesso e gli errori sono facili da commettere, quindi NetApp consiglia vivamente di testare attentamente le operazioni di acquisizione prima di mettere in produzione un sistema di storage. Solo così si può garantire che tutti i percorsi di rete siano configurati correttamente. In un ambiente SAN, verificare lo stato del percorso utilizzando il comando `sanlun lun show -p` o gli strumenti multipathing nativi del sistema operativo per garantire che tutti i percorsi previsti siano disponibili. I sistemi ASA r2 forniscono tutti i percorsi ottimizzati attivi per le LUN e i clienti che utilizzano namespace NVMe dovrebbero affidarsi a strumenti nativi del sistema operativo, poiché i percorsi NVMe non sono coperti da sanlun.

Quando si emette un'acquisizione forzata, è necessario prestare attenzione. Forzare una modifica alla configurazione di archiviazione significa che lo stato del controller proprietario delle unità viene ignorato e il nodo alternativo assume forzatamente il controllo delle unità. L'esecuzione non corretta di un'acquisizione può comportare la perdita o il danneggiamento dei dati, poiché un'acquisizione forzata può eliminare il contenuto di NVMEM. Una volta completata l'acquisizione, la perdita di tali dati implica che i dati memorizzati sulle unità potrebbero tornare a uno stato leggermente più vecchio dal punto di vista del database.

Un'acquisizione forzata con una coppia HA normale dovrebbe essere raramente necessaria. In quasi tutti gli scenari di errore, un nodo si spegne e informa il partner in modo che venga eseguito un failover automatico. Esistono alcuni casi limite, come un rolling failure in cui l'interconnessione tra i nodi viene persa e poi un controller fallisce, in cui è necessario un takeover forzato. In una situazione del genere, il mirroring tra i nodi viene perso prima del guasto del controller, il che significa che il controller ancora in vita non dispone più di una copia delle scritture in corso. L'acquisizione deve quindi essere forzata, il che significa che i dati potrebbero andare persi.

NetApp consiglia di adottare le seguenti precauzioni:

- Fate molta attenzione a non forzare accidentalmente un'acquisizione. Di solito, non dovrebbe essere necessario forzare la modifica, che può causare la perdita di dati.
- Se è necessario un takeover forzato, assicurarsi che le applicazioni siano chiuse, che tutti i file system siano smontati e che i gruppi di volumi del gestore dei volumi logici (LVM) siano disattivati. I gruppi di dischi ASM devono essere smontati.
- In caso di errore a livello di sito durante l'utilizzo di SM-as, il failover automatico non pianificato assistito ONTAP Mediator verrà avviato sul cluster attivo, determinando una breve pausa I/O e quindi le transizioni del database continueranno dal cluster attivo. Per maggiori informazioni, vedere il "Sincronizzazione attiva SnapMirror sui sistemi ASA r2" per i passaggi di configurazione dettagliati.

Configurazione del database con sistemi AFF/ FAS

Dimensioni dei blocchi

ONTAP utilizza internamente dimensioni dei blocchi variabili, il che significa che è possibile configurare i database Oracle con le dimensioni desiderate per i blocchi. Tuttavia, le dimensioni dei blocchi del file system possono influire sulle prestazioni e in alcuni casi una maggiore dimensione dei blocchi di ripristino può migliorare le prestazioni.

Dimensioni dei blocchi di dati

Alcuni sistemi operativi offrono una scelta di dimensioni dei blocchi del file system. Per i file system che supportano i file di dati Oracle, quando si utilizza la compressione le dimensioni del blocco devono essere pari a 8KB KB. Quando la compressione non è necessaria, è possibile utilizzare dimensioni del blocco pari a 8KB K o 4KB K.

Se un file dati viene inserito in un file system con un blocco di 512 byte, è possibile che i file non siano allineati correttamente. Il LUN e il file system potrebbero essere allineati correttamente in base ai consigli di NetApp, ma l'i/o del file non sarebbe allineato correttamente. Un tale disallineamento causerebbe gravi problemi di prestazioni.

I file system che supportano i log di ripristino devono utilizzare una dimensione blocco pari a un multiplo della dimensione del blocco di ripristino. In genere, questo richiede che sia il file system del redo log sia il redo log stesso utilizzino una dimensione del blocco di 512 byte.

Ripristina le dimensioni dei blocchi

A velocità di ripristino molto elevate, è possibile che le dimensioni dei blocchi di 4KB KB funzionino meglio, perché alte velocità di ripristino consentono di eseguire l'i/o in un numero inferiore di operazioni più efficienti. Se le velocità di ripristino sono superiori a 50Mbps KB, valutare la possibilità di testare dimensioni blocco di

4KB KB.

Sono stati identificati alcuni problemi dei clienti con i database che utilizzano i log di redo con dimensioni dei blocchi di 512 byte in un file system con dimensioni dei blocchi di 4KB KB e molte transazioni di dimensioni molto ridotte. L'overhead coinvolto nell'applicazione di modifiche multiple a 512 byte a un singolo blocco di file system da 4KB ha portato a problemi di performance che sono stati risolti modificando il file system in modo da utilizzare dimensioni dei blocchi di 512 byte.

NetApp consiglia di non modificare le dimensioni del blocco di redo se non dietro indicazione di un'organizzazione di assistenza clienti o servizi professionali o se la modifica si basa sulla documentazione ufficiale del prodotto.

db_file_multiblock_read_count

Il `db_file_multiblock_read_count` Parametro controlla il numero massimo di blocchi di database Oracle che Oracle legge come singola operazione durante l'i/o sequenziale

Questo parametro non influisce tuttavia sul numero di blocchi letti da Oracle durante qualsiasi e in tutte le operazioni di lettura, né sull'i/o casuale Ciò influisce solo sulle dimensioni del blocco degli i/o sequenziali.

Oracle consiglia all'utente di lasciare il parametro non impostato. In questo modo, il software del database può impostare automaticamente il valore ottimale. Questo generalmente significa che questo parametro è impostato su un valore che fornisce una dimensione i/o di 1MB. Ad esempio, una lettura 1MB di 8KB blocchi richiederebbe la lettura di 128 blocchi e il valore predefinito per questo parametro sarebbe 128.

La maggior parte dei problemi di prestazioni del database osservati da NetApp presso le sedi dei clienti implica un'impostazione errata per questo parametro. Ci sono motivi validi per modificare questo valore con le versioni 8 e 9 di Oracle. Di conseguenza, il parametro potrebbe essere inconsapevolmente presente in `init.ora` Perché il database è stato aggiornato in posizione a Oracle 10 e versioni successive. Un'impostazione legacy di 8 o 16, rispetto a un valore predefinito di 128, danneggia significativamente le prestazioni i/o sequenziali.

NetApp recommended impostando `db_file_multiblock_read_count` il parametro non deve essere presente in `init.ora` file. NetApp non ha mai riscontrato una situazione in cui la modifica di questo parametro ha migliorato le prestazioni, ma in molti casi ha causato evidenti danni al throughput i/o sequenziale.

filesystemio_options

Parametro di inizializzazione Oracle `filesystemio_options` Controlla l'utilizzo dell'i/o asincrono e diretto

Contrariamente a quanto si crede, l'i/o asincrono e diretto non si escludono a vicenda. NetApp ha osservato che questo parametro è spesso configurato in modo non corretto negli ambienti dei clienti e che questa errata configurazione è direttamente responsabile di molti problemi di prestazioni.

L'i/o asincrono significa che le operazioni i/o di Oracle possono essere parallelizzate. Prima della disponibilità di i/o asincrono su vari sistemi operativi, gli utenti hanno configurato numerosi processi dbwriter e modificato la configurazione del processo del server. Con l'i/o asincrono, il sistema operativo stesso esegue i/o per conto del software di database in modo altamente efficiente e parallelo. La procedura non pone i dati a rischio e le operazioni critiche, come il logging di redo di Oracle, vengono comunque eseguite in maniera sincrona.

L'i/o diretto ignora la cache buffer del sistema operativo. L'i/o su un sistema UNIX scorre normalmente attraverso la cache del buffer del sistema operativo. Ciò è utile per le applicazioni che non mantengono una cache interna, ma Oracle dispone di una propria cache buffer all'interno di SGA. In quasi tutti i casi, è meglio abilitare l'i/o diretto e allocare la RAM del server all'SGA piuttosto che affidarsi alla cache del buffer del sistema operativo. Oracle SGA utilizza la memoria in modo più efficiente. Inoltre, quando l'i/o fluisce attraverso il buffer del sistema operativo, è soggetto a un'ulteriore elaborazione, che aumenta le latenze. L'aumento delle latenze è particolarmente percepibile con un elevato i/o in scrittura quando una bassa latenza è un requisito critico.

Le opzioni per `filesystemio_options` sono:

- **Async.** Oracle invia le richieste di i/o al sistema operativo per l'elaborazione. Questo processo consente a Oracle di eseguire altri lavori anziché attendere il completamento dell'i/o e quindi aumentare la parallelizzazione i/O.
- **Directio.** Oracle esegue l'i/o direttamente sui file fisici piuttosto che instradare l'i/o attraverso la cache del sistema operativo host.
- **None.** Oracle utilizza i/o sincroni e bufferizzati In questa configurazione, la scelta tra processi server condivisi e dedicati e il numero di dbwriter è più importante.
- **Setall.** Oracle utilizza i/o sia asincrono che diretto In quasi tutti i casi, l'uso di `setall` è ottimale.

Il `filesystemio_options` Parametro non ha effetto negli ambienti DNFS e ASM. L'utilizzo di DNFS o ASM comporta automaticamente l'utilizzo dell'i/o asincrono e diretto

In passato alcuni clienti hanno riscontrato problemi di i/o asincrono, in particolare con le precedenti versioni di Red Hat Enterprise Linux 4 (RHEL4). Alcuni consigli aggiornati su Internet suggeriscono comunque di evitare l'i/o asincrono a causa di informazioni non aggiornate. L'i/o asincrono è stabile su tutti i sistemi operativi correnti. Non c'è motivo di disabilitarlo, in assenza di un bug noto con il sistema operativo.

Se un database utilizza l'i/o con buffer, un passaggio all'i/o diretto potrebbe anche richiedere una modifica delle dimensioni SGA. La disattivazione dell'i/o con buffer elimina i vantaggi prestazionali che la cache del sistema operativo host fornisce al database. L'aggiunta di RAM alla SGA risolve questo problema. Il risultato netto dovrebbe essere un miglioramento delle performance di i/O.

Sebbene sia quasi sempre meglio utilizzare la RAM per Oracle SGA piuttosto che per il caching del buffer del sistema operativo, potrebbe essere impossibile determinare il valore migliore. Ad esempio, potrebbe essere preferibile utilizzare i/o con buffer di dimensioni SGA molto ridotte su un server di database con molte istanze Oracle attive in modo intermittente. Questa disposizione consente l'utilizzo flessibile della RAM disponibile rimanente sul sistema operativo da parte di tutte le istanze di database in esecuzione. Si tratta di una situazione molto insolita, ma è stata osservata presso alcune sedi dei clienti.

NetApp recommended `filesystemio_options` a. `setall`, Ma essere consapevoli che in alcune circostanze la perdita della cache del buffer host potrebbe richiedere un aumento nella SGA di Oracle.

Timeout RAC

Oracle RAC è un prodotto clusterware con diversi tipi di processi heartbeat interni che monitorano lo stato del cluster.

Le informazioni contenute in "errore di montaggio" La sezione contiene informazioni critiche per gli ambienti Oracle RAC che utilizzano lo storage di rete e, in molti casi, è necessario modificare le impostazioni predefinite di Oracle RAC per garantire che il cluster RAC sopravviva alle modifiche del percorso di rete e alle operazioni di failover/switchover dello storage.

disktimeout

Il parametro RAC relativo allo storage primario è `disktimeout`. Questo parametro controlla la soglia entro la quale l'i/o del file di voting deve essere completato. Se il `disktimeout` Il parametro viene superato, quindi il nodo RAC viene eliminato dal cluster. Il valore predefinito per questo parametro è 200. Questo valore dovrebbe essere sufficiente per le procedure standard di takeover e giveback dello storage.

NetApp consiglia di eseguire il test approfondito delle configurazioni RAC prima di metterle in produzione, perché molti fattori influiscono su un takeover o un giveback. Oltre al tempo richiesto per il completamento del failover dello storage, è necessario ulteriore tempo affinché le modifiche LACP (link Aggregation Control Protocol) vengano propagate. Inoltre, il software multipathing SAN deve rilevare un timeout i/o e riprovare su un percorso alternativo. Se un database è estremamente attivo, è necessario accodare e rieseguire una grande quantità di i/o prima di elaborare l'i/o del disco di voting.

Nel caso in cui non sia possibile eseguire un takeover o un giveback effettivo dello storage, l'effetto può essere simulato con test di pull dei cavi sul server di database.

NetApp consiglia quanto segue:

- Lasciando il `disktimeout` parametro al valore predefinito di 200.
- Verificare sempre accuratamente la configurazione di un RAC.

errore di montaggio

Il `misscount` In genere, il parametro influisce solo sull'heartbeat di rete tra i nodi RAC. Il valore predefinito è 30 secondi. Se i binari della griglia si trovano su un array di storage o l'unità di avvio del sistema operativo non è locale, questo parametro potrebbe diventare importante. Ciò comprende host con unità di boot ubicate su una SAN FC, sistemi operativi basati su NFS e unità di boot ubicate in datastore di virtualizzazione, come un file VMDK.

Se l'accesso a un'unità di boot viene interrotto da un takeover o un giveback dello storage, è possibile che la posizione binaria della griglia o l'intero sistema operativo si blocchi temporaneamente. Il tempo necessario affinché ONTAP completi l'operazione di storage e affinché il sistema operativo modifichi i percorsi e riprenda l'i/o potrebbe superare l' `misscount` soglia. Di conseguenza, un nodo viene eliminato immediatamente dopo il ripristino della connettività al LUN di avvio o ai binari della griglia. Nella maggior parte dei casi, l'eviction e il successivo riavvio si verificano senza messaggi di registrazione per indicare il motivo del riavvio. Non tutte le configurazioni sono interessate dal problema, pertanto è possibile testare host basati su boot SAN, NFS o datastore in un ambiente RAC in modo che RAC rimanga stabile in caso di interruzione della comunicazione con il disco di avvio.

Nel caso di unità di avvio non locali o di un hosting di file system non locale `grid` binari, il `misscount` deve essere modificato per corrispondere `disktimeout`. Se questo parametro viene modificato, eseguire ulteriori test per identificare anche eventuali effetti sul comportamento RAC, come il tempo di failover dei nodi.

NetApp consiglia quanto segue:

- Lasciare la `misscount` parametro al valore predefinito di 30 a meno che non si verifichi una delle seguenti condizioni:
  - `grid` I file binari sono collocati in un disco collegato in rete, inclusi dischi NFS, iSCSI, FC e basati su datastore.
  - Il sistema operativo viene avviato SAN.
- In questi casi, valutare l'effetto delle interruzioni di rete che influiscono sull'accesso al sistema operativo o. `GRID_HOME` file system. In alcuni casi, tali interruzioni causano lo stallo dei daemon Oracle RAC, che può portare a un `misscount`timeout e sfratto basati su -. Il valore predefinito del timeout è 27 secondi, ovvero il valore di `misscount` meno `reboottime`. In questi casi, aumentare `misscount` a 200 per la corrispondenza `disktimeout`.

Configurazione del database con sistemi ASA r2

Dimensioni dei blocchi

ONTAP utilizza internamente una dimensione di blocco variabile, il che significa che i database Oracle possono essere configurati con qualsiasi dimensione di blocco desiderata. Tuttavia, le dimensioni dei blocchi del file system possono influire sulle prestazioni e, in alcuni casi, una dimensione maggiore del blocco redo può migliorare le prestazioni.

ASA r2 non introduce alcuna modifica alle raccomandazioni sulle dimensioni dei blocchi Oracle rispetto ai sistemi AFF/ FAS . Il comportamento ONTAP rimane coerente su tutte le piattaforme.

Dimensioni dei blocchi di dati

Alcuni sistemi operativi offrono una scelta di dimensioni dei blocchi del file system. Per i file system che supportano i file di dati Oracle, quando si utilizza la compressione le dimensioni del blocco devono essere pari a 8KB KB. Quando la compressione non è necessaria, è possibile utilizzare dimensioni del blocco pari a 8KB K o 4KB K.

Se un file dati viene inserito in un file system con un blocco di 512 byte, è possibile che i file non siano allineati correttamente. Il LUN e il file system potrebbero essere allineati correttamente in base ai consigli di NetApp, ma l'i/o del file non sarebbe allineato correttamente. Un tale disallineamento causerebbe gravi problemi di prestazioni.

Ripristina le dimensioni dei blocchi

I file system che supportano i log di ripristino devono utilizzare una dimensione blocco pari a un multiplo della dimensione del blocco di ripristino. In genere, questo richiede che sia il file system del redo log sia il redo log stesso utilizzino una dimensione del blocco di 512 byte.

A velocità di ripristino molto elevate, è possibile che le dimensioni dei blocchi di 4KB KB funzionino meglio, perché alte velocità di ripristino consentono di eseguire l'i/o in un numero inferiore di operazioni più efficienti. Se le velocità di ripristino sono superiori a 50Mbps KB, valutare la possibilità di testare dimensioni blocco di 4KB KB.

Sono stati identificati alcuni problemi dei clienti con i database che utilizzano i log di redo con dimensioni dei blocchi di 512 byte in un file system con dimensioni dei blocchi di 4KB KB e molte transazioni di dimensioni molto ridotte. L'overhead coinvolto nell'applicazione di modifiche multiple a 512 byte a un singolo blocco di file system da 4KB ha portato a problemi di performance che sono stati risolti modificando il file system in modo da utilizzare dimensioni dei blocchi di 512 byte.

NetApp consiglia di non modificare le dimensioni del blocco di redo se non dietro indicazione di un'organizzazione di assistenza clienti o servizi professionali o se la modifica si basa sulla documentazione ufficiale del prodotto.

db_file_multiblock_read_count

Il `db_file_multiblock_read_count` Parametro controlla il numero massimo di blocchi di database Oracle che Oracle legge come singola operazione durante l'i/o sequenziale

Non ci sono cambiamenti nelle raccomandazioni rispetto ai sistemi AFF/ FAS . Il comportamento ONTAP e le best practice di Oracle rimangono identici sulle piattaforme ASA r2, AFF e FAS .

Questo parametro non influisce tuttavia sul numero di blocchi letti da Oracle durante qualsiasi e in tutte le operazioni di lettura, né sull'i/o casuale Ciò influisce solo sulle dimensioni del blocco degli i/o sequenziali.

Oracle consiglia all'utente di lasciare il parametro non impostato. In questo modo, il software del database può impostare automaticamente il valore ottimale. Questo generalmente significa che questo parametro è impostato su un valore che fornisce una dimensione i/o di 1MB. Ad esempio, una lettura 1MB di 8KB blocchi richiederebbe la lettura di 128 blocchi e il valore predefinito per questo parametro sarebbe 128.

La maggior parte dei problemi di prestazioni del database osservati da NetApp presso le sedi dei clienti implica un'impostazione errata per questo parametro. Ci sono motivi validi per modificare questo valore con le versioni 8 e 9 di Oracle. Di conseguenza, il parametro potrebbe essere inconsapevolmente presente in `init.ora` Perché il database è stato aggiornato in posizione a Oracle 10 e versioni successive. Un'impostazione legacy di 8 o 16, rispetto a un valore predefinito di 128, danneggia significativamente le prestazioni i/o sequenziali.

NetApp recommended impostando `db_file_multiblock_read_count` il parametro non deve essere presente in `init.ora` file. NetApp non ha mai riscontrato una situazione in cui la modifica di questo parametro ha migliorato le prestazioni, ma in molti casi ha causato evidenti danni al throughput i/o sequenziale.

filesystemio_options

Parametro di inizializzazione Oracle `filesystemio_options` Controlla l'utilizzo dell'i/o asincrono e diretto

Il comportamento e le raccomandazioni per filesystemio_options su ASA r2 sono identici ai sistemi AFF/ FAS perché il parametro è specifico di Oracle e non dipende dalla piattaforma di archiviazione. ASA r2 utilizza ONTAP come AFF/ FAS, quindi si applicano le stesse best practice.

Contrariamente a quanto si crede, l'i/o asincrono e diretto non si escludono a vicenda. NetApp ha osservato che questo parametro è spesso configurato in modo non corretto negli ambienti dei clienti e che questa errata configurazione è direttamente responsabile di molti problemi di prestazioni.

L'i/o asincrono significa che le operazioni i/o di Oracle possono essere parallelizzate. Prima della disponibilità

di i/o asincrono su vari sistemi operativi, gli utenti hanno configurato numerosi processi dbwriter e modificato la configurazione del processo del server. Con l'i/o asincrono, il sistema operativo stesso esegue i/o per conto del software di database in modo altamente efficiente e parallelo. La procedura non pone i dati a rischio e le operazioni critiche, come il logging di redo di Oracle, vengono comunque eseguite in maniera sincrona.

L'i/o diretto ignora la cache buffer del sistema operativo. L'i/o su un sistema UNIX scorre normalmente attraverso la cache del buffer del sistema operativo. Ciò è utile per le applicazioni che non mantengono una cache interna, ma Oracle dispone di una propria cache buffer all'interno di SGA. In quasi tutti i casi, è meglio abilitare l'i/o diretto e allocare la RAM del server all'SGA piuttosto che affidarsi alla cache del buffer del sistema operativo. Oracle SGA utilizza la memoria in modo più efficiente. Inoltre, quando l'i/o fluisce attraverso il buffer del sistema operativo, è soggetto a un'ulteriore elaborazione, che aumenta le latenze. L'aumento delle latenze è particolarmente percepibile con un elevato i/o in scrittura quando una bassa latenza è un requisito critico.

Le opzioni per `filesystemio_options` sono:

- **Async.** Oracle invia le richieste di i/o al sistema operativo per l'elaborazione. Questo processo consente a Oracle di eseguire altri lavori anziché attendere il completamento dell'i/o e quindi aumentare la parallelizzazione i/O.
- **Directio.** Oracle esegue l'i/o direttamente sui file fisici piuttosto che instradare l'i/o attraverso la cache del sistema operativo host.
- **None.** Oracle utilizza i/o sincroni e bufferizzati In questa configurazione, la scelta tra processi server condivisi e dedicati e il numero di dbwriter è più importante.
- **Setall.** Oracle utilizza i/o sia asincrono che diretto In quasi tutti i casi, l'uso di `setall` è ottimale.

Negli ambienti ASM, Oracle utilizza automaticamente I/O diretto e I/O asincrono per i dischi gestiti da ASM, quindi `filesystemio_options` non ha alcun effetto sui gruppi di dischi ASM. Per le distribuzioni non ASM (ad esempio, file system su LUN SAN), impostare: `filesystemio_options = setall`. Ciò consente sia l'I/O asincrono che quello diretto per prestazioni ottimali.

Alcuni vecchi sistemi operativi presentavano problemi con l'I/O asincrono, il che ha portato a consigli obsoleti che suggerivano di evitarlo. Tuttavia, l'I/O asincrono è stabile e pienamente supportato su tutti i sistemi operativi attuali. Non c'è motivo di disattivarlo a meno che non venga identificato un bug specifico del sistema operativo.

Se un database utilizza l'i/o con buffer, un passaggio all'i/o diretto potrebbe anche richiedere una modifica delle dimensioni SGA. La disattivazione dell'i/o con buffer elimina i vantaggi prestazionali che la cache del sistema operativo host fornisce al database. L'aggiunta di RAM alla SGA risolve questo problema. Il risultato netto dovrebbe essere un miglioramento delle performance di i/O.

Sebbene sia quasi sempre meglio utilizzare la RAM per Oracle SGA piuttosto che per il caching del buffer del sistema operativo, potrebbe essere impossibile determinare il valore migliore. Ad esempio, potrebbe essere preferibile utilizzare i/o con buffer di dimensioni SGA molto ridotte su un server di database con molte istanze Oracle attive in modo intermittente. Questa disposizione consente l'utilizzo flessibile della RAM disponibile rimanente sul sistema operativo da parte di tutte le istanze di database in esecuzione. Si tratta di una situazione molto insolita, ma è stata osservata presso alcune sedi dei clienti.

* NetApp consiglia* l'impostazione `filesystemio_options` A `setall`, ma tieni presente che in alcune circostanze la perdita della cache buffer dell'host potrebbe richiedere un aumento dell'SGA di Oracle. I sistemi ASA r2 sono ottimizzati per carichi di lavoro SAN con bassa latenza, pertanto l'utilizzo di setall si allinea perfettamente con la progettazione di ASA per distribuzioni Oracle ad alte prestazioni.

Timeout RAC

Oracle RAC è un prodotto clusterware con diversi tipi di processi heartbeat interni che monitorano lo stato del cluster.

I sistemi ASA r2 utilizzano ONTAP proprio come AFF/ FAS, quindi gli stessi principi si applicano ai parametri di timeout di Oracle RAC. Non ci sono modifiche specifiche ASA alle raccomandazioni disktimeout o misscount. Tuttavia, ASA r2 è ottimizzato per carichi di lavoro SAN e failover a bassa latenza, il che rende queste best practice ancora più critiche.

Le informazioni nel "errore di montaggio" La sezione include informazioni critiche per gli ambienti Oracle RAC che utilizzano storage in rete e, in molti casi, sarà necessario modificare le impostazioni predefinite di Oracle RAC per garantire che il cluster RAC sopravviva alle modifiche del percorso di rete e alle operazioni di failover dello storage.

disktimeout

Il parametro RAC relativo allo storage primario è `disktimeout`. Questo parametro controlla la soglia entro la quale l'i/o del file di voting deve essere completato. Se il `disktimeout` Il parametro viene superato, quindi il nodo RAC viene eliminato dal cluster. Il valore predefinito per questo parametro è 200. Questo valore dovrebbe essere sufficiente per le procedure standard di takeover e giveback dello storage.

NetApp consiglia di eseguire il test approfondito delle configurazioni RAC prima di metterle in produzione, perché molti fattori influiscono su un takeover o un giveback. Oltre al tempo richiesto per il completamento del failover dello storage, è necessario ulteriore tempo affinché le modifiche LACP (link Aggregation Control Protocol) vengano propagate. Inoltre, il software multipathing SAN deve rilevare un timeout i/o e riprovare su un percorso alternativo. Se un database è estremamente attivo, è necessario accodare e rieseguire una grande quantità di i/o prima di elaborare l'i/o del disco di voting.

Nel caso in cui non sia possibile eseguire un takeover o un giveback effettivo dello storage, l'effetto può essere simulato con test di pull dei cavi sul server di database.

NetApp consiglia quanto segue:

- Lasciando il `disktimeout` parametro al valore predefinito di 200.
- Verificare sempre accuratamente la configurazione di un RAC.

errore di montaggio

Il `misscount` In genere, il parametro influisce solo sull'heartbeat di rete tra i nodi RAC. Il valore predefinito è 30 secondi. Se i binari della griglia si trovano su un array di storage o l'unità di avvio del sistema operativo non è locale, questo parametro potrebbe diventare importante. Ciò comprende host con unità di boot ubicate su una SAN FC, sistemi operativi basati su NFS e unità di boot ubicate in datastore di virtualizzazione, come un file VMDK.

Se l'accesso a un'unità di boot viene interrotto da un takeover o un giveback dello storage, è possibile che la posizione binaria della griglia o l'intero sistema operativo si blocchi temporaneamente. Il tempo necessario affinché ONTAP completi l'operazione di storage e affinché il sistema operativo modifichi i percorsi e riprenda l'i/o potrebbe superare l' `misscount` soglia. Di conseguenza, un nodo viene eliminato immediatamente dopo il ripristino della connettività al LUN di avvio o ai binari della griglia. Nella maggior parte dei casi, l'eviction e il successivo riavvio si verificano senza messaggi di registrazione per indicare il motivo del riavvio. Non tutte le configurazioni sono interessate dal problema, pertanto è possibile testare host basati su boot SAN, NFS o

datastore in un ambiente RAC in modo che RAC rimanga stabile in caso di interruzione della comunicazione con il disco di avvio.

Nel caso di unità di avvio non locali o di un hosting di file system non locale `grid` binari, il `misscount` deve essere modificato per corrispondere `disktimeout`. Se questo parametro viene modificato, eseguire ulteriori test per identificare anche eventuali effetti sul comportamento RAC, come il tempo di failover dei nodi.

NetApp consiglia quanto segue:

- Lasciare la `misscount` parametro al valore predefinito di 30 a meno che non si verifichi una delle seguenti condizioni:
  - `grid` i file binari si trovano su un'unità collegata alla rete, tra cui unità iSCSI, FC e basate su datastore.
  - Il sistema operativo viene avviato SAN.
- In questi casi, valutare l'effetto delle interruzioni di rete che influiscono sull'accesso al sistema operativo o. `GRID_HOME` file system. In alcuni casi, tali interruzioni causano lo stallo dei daemon Oracle RAC, che può portare a un `misscount`timeout e sfratto basati su -. Il valore predefinito del timeout è 27 secondi, ovvero il valore di `misscount` meno `reboottime`. In questi casi, aumentare `misscount` a 200 per la corrispondenza `disktimeout`.

- Il design ottimizzato per SAN di ASA r2 riduce la latenza del failover, ma i timeout devono comunque essere regolati per l'avvio in rete o per i binari della griglia.
- Per configurazioni RAC estese o attive-attive (ad esempio, SnapMirror ActiveSync), la regolazione del timeout rimane essenziale per le architetture zero-RPO.

Configurazione host con sistemi AFF/ FAS

AIX

Argomenti di configurazione per database Oracle su IBM AIX con ONTAP.

I/o simultanei

Per ottenere prestazioni ottimali su IBM AIX è necessario utilizzare l'i/o simultaneo Senza i/o simultaneo, è probabile che le limitazioni delle prestazioni siano dovute al fatto che AIX esegue i/o atomico serializzato, che comporta un overhead significativo.

In origine, NetApp ha consigliato di utilizzare `cio` Opzione di montaggio per forzare l'uso di i/o simultanei sul file system, ma questo processo ha avuto degli inconvenienti e non è più necessario. Dall'introduzione di AIX 5,2 e Oracle 10gR1, Oracle su AIX può aprire singoli file per i/o simultanei, anziché forzare i/o simultanei sull'intero file system.

Il metodo migliore per abilitare l'i/o simultaneo è impostare `init.ora` parametro `filesystemio_options` a. `setall`. In questo modo, Oracle può aprire file specifici da utilizzare con i/o simultanei

Utilizzo di `cio` Come opzione di montaggio, l'utilizzo di i/o simultanei può avere conseguenze negative. Ad esempio, forzando i/o simultanei si disabilita la lettura dei file system, che può danneggiare le prestazioni dell'i/o al di fuori del software del database Oracle, come la copia dei file e l'esecuzione di backup su nastro. Inoltre, prodotti come Oracle GoldenGate e SAP BR*Tools non sono compatibili con l'uso di `cio` Montare

l'opzione con alcune versioni di Oracle.

NetApp consiglia quanto segue:

- Non utilizzare `cio` opzione di montaggio a livello di file system. Abilitare invece l'i/o simultaneo tramite l'utilizzo di `filesystemio_options=setall`.
- Utilizzare solo l' `cio` l'opzione di montaggio dovrebbe essere impostata se non è possibile `filesystemio_options=setall`.

Opzioni di montaggio NFS AIX

Nella tabella seguente sono elencate le opzioni di montaggio NFS AIX per i database Oracle a istanza singola.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
File di controllo File di dati Registri di ripristino	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, intr`

Nella tabella seguente sono elencate le opzioni di montaggio NFS AIX per RAC.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
File di controllo File di dati Registri di ripristino	`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`
Dedicato `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Condiviso `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

L'aggiunta fa la differenza principale tra le opzioni di montaggio RAC e a istanza singola `noac` alle opzioni di montaggio. Questa aggiunta ha l'effetto di disabilitare la cache del sistema operativo host, consentendo a tutte le istanze nel cluster RAC di avere una visione coerente dello stato dei dati.

Anche se si utilizza il `cio` montare l'opzione e. `init.ora` parametro `filesystemio_options=setall` ha lo stesso effetto di disabilitare la cache dell'host, è comunque necessario utilizzare `noac`. `noac` è obbligatorio

per condiviso `ORACLE_HOME` Implementazioni per facilitare la coerenza di file quali file di password Oracle e. `spfile` file di parametri. Se ogni istanza di un cluster RAC dispone di un'istanza dedicata `ORACLE_HOME`, questo parametro non è necessario.

Opzioni di montaggio di AIX jfs/JFS2

Nella tabella seguente sono elencate le opzioni di montaggio di AIX jfs/JFS2.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	Valori predefiniti
File di controllo File di dati Registri di ripristino	Valori predefiniti
ORACLE_HOME	Valori predefiniti

Prima di utilizzare AIX `hdisk` i dispositivi in qualsiasi ambiente, inclusi i database, controllano il parametro `queue_depth`. Questo parametro non è la profondità della coda HBA, bensì la profondità della coda SCSI dell'individuo `hdisk device. Depending on how the LUNs are configured, the value for `queue_depth` potrebbe essere troppo basso per garantire buone prestazioni. I test hanno dimostrato che il valore ottimale è 64.

HP-UX

Argomenti di configurazione per database Oracle su HP-UX con ONTAP.

Opzioni di montaggio NFS HP-UX

Nella tabella seguente sono elencate le opzioni di montaggio NFS HP-UX per una singola istanza.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
File di controllo File di dati Registri di ripristino	`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`

Nella tabella seguente sono elencate le opzioni di montaggio NFS HP-UX per RAC.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac,suid`

Tipo di file	Opzioni di montaggio
File di controllo File di dati Registri di ripristino	`rw, bg,hard, [vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio,suid`
CRS/votazione	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac, forcedirectio,suid`
Dedicato `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Condiviso `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,suid`

L'aggiunta fa la differenza principale tra le opzioni di montaggio RAC e a istanza singola `noac` e. `forcedirectio` alle opzioni di montaggio. Questa aggiunta ha l'effetto di disabilitare il caching del sistema operativo host, consentendo a tutte le istanze nel cluster RAC di avere una visione coerente dello stato dei dati. Anche se si utilizza il `init.ora` parametro `filesystemio_options=setall` ha lo stesso effetto di disabilitare la cache dell'host, è comunque necessario utilizzare `noac` e. `forcedirectio`.

Il motivo `noac` è obbligatorio per condiviso `ORACLE_HOME` Le distribuzioni consentono di semplificare la coerenza di file quali file di password Oracle e file spfile. Se ogni istanza di un cluster RAC dispone di un'istanza dedicata `ORACLE_HOME`, questo parametro non è richiesto.

Opzioni di montaggio VxFS HP-UX

Utilizzare le seguenti opzioni di montaggio per i file system che ospitano file binari Oracle:

```
delaylog,nodatainlog
```

Utilizzare le seguenti opzioni di montaggio per i file system contenenti file di dati, log di ripristino, log di archivio e file di controllo in cui la versione di HP-UX non supporta i/o simultanei:

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

Quando l'i/o simultaneo è supportato (VxFS 5.0.1 e versioni successive o con ServiceGuard Storage Management Suite), utilizzare queste opzioni di montaggio per i file system contenenti file di dati, log di ripristino, log di archivio e file di controllo:

```
delaylog,cio
```

Il parametro `db_file_multiblock_read_count` È particolarmente critico negli ambienti VxFS. Oracle consiglia di non impostare questo parametro in Oracle 10g R1 e versioni successive, a meno che non sia diversamente specificato. L'impostazione predefinita con dimensioni blocco Oracle 8KB è 128 KB. Se il valore di questo parametro è forzato a 16 o inferiore, rimuovere l' `convosync=direct` Montare l'opzione perché può danneggiare le prestazioni i/o sequenziali. Questa operazione danneggia altri aspetti delle prestazioni e deve essere eseguita solo se il valore di `db_file_multiblock_read_count` deve essere modificato dal valore predefinito.

Linux

Argomenti di configurazione specifici del sistema operativo Linux.

Tabelle degli slot TCP per Linux NFSv3

Le tabelle degli slot TCP sono l'equivalente di NFSv3 della profondità della coda degli HBA (host Bus Adapter). Queste tabelle controllano il numero di operazioni NFS che possono essere in sospeso in qualsiasi momento. Il valore predefinito è di solito 16, che è troppo basso per ottenere prestazioni ottimali. Il problema opposto si verifica sui kernel Linux più recenti, che possono aumentare automaticamente il limite della tabella degli slot TCP a un livello che satura il server NFS con le richieste.

Per prestazioni ottimali e per evitare problemi di prestazioni, regolare i parametri del kernel che controllano le tabelle degli slot TCP.

Eseguire `sysctl -a | grep tcp.*.slot_table` e osservare i seguenti parametri:

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

Tutti i sistemi Linux dovrebbero includere `sunrpc.tcp_slot_table_entries`, ma solo alcuni includono `sunrpc.tcp_max_slot_table_entries`. Entrambi devono essere impostati su 128.

La mancata impostazione di questi parametri può avere effetti significativi sulle prestazioni. In alcuni casi, le prestazioni sono limitate poiché il sistema operativo linux non fornisce i/o sufficienti In altri casi, le latenze i/o aumentano quando il sistema operativo linux tenta di emettere più i/o di quanto possa essere gestito.

Opzioni di montaggio NFS Linux

Nella tabella seguente sono elencate le opzioni di montaggio NFS Linux per una singola istanza.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
File di controllo File di dati Registri di ripristino	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

Tipo di file	Opzioni di montaggio
ORACLE_HOME	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

Nella tabella seguente sono elencate le opzioni di montaggio NFS Linux per RAC.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, actimeo=0`
File di controllo File di dati Registri di ripristino	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`
CRS/votazione	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,actimeo=0`
Dedicato `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Condiviso `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`

L'aggiunta fa la differenza principale tra le opzioni di montaggio RAC e a istanza singola `actimeo=0` alle opzioni di montaggio. Questa aggiunta ha l'effetto di disabilitare il caching del sistema operativo host, consentendo a tutte le istanze nel cluster RAC di avere una visione coerente dello stato dei dati. Anche se si utilizza il `init.ora` parametro `filesystemio_options=setall` ha lo stesso effetto di disabilitare la cache dell'host, è comunque necessario utilizzare `actimeo=0`.

Il motivo `actimeo=0` è obbligatorio per condiviso `ORACLE_HOME` Le distribuzioni consentono di semplificare la coerenza di file quali file di password e file spfile di Oracle. Se ogni istanza di un cluster RAC dispone di un'istanza dedicata `ORACLE_HOME`, questo parametro non è necessario.

In genere, i file non di database devono essere montati con le stesse opzioni utilizzate per i file di dati a singola istanza, sebbene applicazioni specifiche possano avere requisiti diversi. Evitare le opzioni di montaggio `noac` e. `actimeo=0` se possibile perché queste opzioni disabilitano la lettura e il buffering a livello di file system. Ciò può causare gravi problemi di prestazioni per processi quali l'estrazione, la traduzione e il caricamento.

ACCESSO e GETATTR

Alcuni clienti hanno notato che un livello estremamente elevato di altri IOPS, come ACCESSO e GETATTR, può dominare i propri workload. In casi estremi, operazioni come letture e scritture possono arrivare fino al 10% del totale. Si tratta di un comportamento normale con qualsiasi database che include l'uso di `actimeo=0` e/o. `noac` Su Linux perché queste opzioni fanno sì che il sistema operativo Linux ricarichi costantemente i metadati dei file dal sistema di archiviazione. Operazioni quali ACCESS e GETATTR sono operazioni a basso impatto gestite dalla cache ONTAP in un ambiente di database. Non dovrebbero essere considerati IOPS autentici, come le letture e le scritture, che creano una vera domanda sui sistemi storage. Tuttavia, questi altri IOPS creano un certo carico, specialmente negli ambienti RAC. Per risolvere questo problema, abilitare DNFS, che ignora la cache buffer del sistema operativo ed evita queste operazioni non necessarie relative ai

metadati.

Linux Direct NFS

Un'opzione di montaggio aggiuntiva, denominata `nosharecache`, È necessario quando (a) DNFS è abilitato e (b) un volume sorgente è montato più di una volta su un singolo server (c) con un mount NFS nidificato. Questa configurazione si osserva principalmente in ambienti che supportano applicazioni SAP. Ad esempio, un singolo volume di un sistema NetApp può avere una directory situata in `/vol/oracle/base` e un secondo a. `/vol/oracle/home`. Se `/vol/oracle/base` è montato su `/oracle` e. `/vol/oracle/home` è montato su `/oracle/home`, Il risultato sono montaggi NFS nidificati che hanno origine sulla stessa fonte.

Il sistema operativo è in grado di rilevare che `/oracle` e `/oracle/home` si trovano sullo stesso volume, che è lo stesso file system di origine. Il sistema operativo utilizza quindi lo stesso handle di dispositivo per l'accesso ai dati. In questo modo si migliora l'uso della cache del sistema operativo e di alcune altre operazioni, ma interferisce con DNFS. Se DNFS deve accedere a un file, come ad esempio `spfile` , on , `/oracle/home` potrebbe erroneamente tentare di utilizzare il percorso errato dei dati. Il risultato è un'operazione i/o non riuscita. In queste configurazioni, Aggiungi l' `nosharecache`opzione di montaggio a qualsiasi file system NFS che condivide un volume di origine con un altro file system NFS su quell'host. In questo modo, il sistema operativo Linux assegna un handle di dispositivo indipendente al file system.

Linux Direct NFS e Oracle RAC

L'uso di DNFS offre speciali vantaggi in termini di prestazioni per Oracle RAC sul sistema operativo Linux, poiché Linux non dispone di un metodo per forzare l'i/o diretto, necessario con RAC per la coerenza tra i nodi. Come soluzione, Linux richiede l'uso di `actimeo=0` Opzione di montaggio, che fa sì che i dati dei file scadano immediatamente dalla cache del sistema operativo. Questa opzione a sua volta obbliga il client NFS Linux a rileggere costantemente i dati degli attributi, danneggiando la latenza e aumentando il carico sullo storage controller.

Abilitando DNFS si ignora il client NFS dell'host ed evita questo danno. Diversi clienti hanno segnalato significativi miglioramenti delle performance sui cluster RAC e una significativa riduzione del carico ONTAP (soprattutto in relazione ad altri IOPS) quando si attiva DNFS.

Linux Direct NFS e file oranfstab

Quando si utilizza DNFS su Linux con l'opzione multipathing, è necessario utilizzare più sottoreti. Su altri sistemi operativi, è possibile stabilire più canali DNFS utilizzando `LOCAL` e. `DONTROUTE` Opzioni per configurare più canali DNFS su una singola subnet. Tuttavia, questo non funziona correttamente su Linux e possono verificarsi problemi di prestazioni imprevisti. Con Linux, ogni NIC utilizzata per il traffico DNFS deve trovarsi su una subnet diversa.

Utilità di pianificazione i/O.

Il kernel Linux permette un controllo di basso livello sul modo in cui l'i/o blocca i dispositivi è programmato. Le impostazioni predefinite su varie distribuzioni di Linux variano notevolmente. I test dimostrano che la scadenza di solito offre i migliori risultati, ma a volte NOOP è stato leggermente migliore. La differenza di prestazioni è minima, ma è necessario verificare entrambe le opzioni se è necessario estrarre le massime prestazioni possibili da una configurazione di database. CFQ è l'impostazione predefinita in molte configurazioni e ha dimostrato di avere problemi significativi di prestazioni con i carichi di lavoro del database.

Per istruzioni sulla configurazione dello scheduler i/o, consultare la documentazione del fornitore di Linux pertinente.

Multipathing

Alcuni clienti hanno riscontrato arresti anomali durante l'interruzione della rete perché il daemon multipath non era in esecuzione sul proprio sistema. Nelle versioni recenti di Linux, il processo di installazione del sistema operativo e del demone multipathing potrebbero lasciare questi sistemi operativi vulnerabili a questo problema. I pacchetti sono installati correttamente, ma non sono configurati per l'avvio automatico dopo un riavvio.

Ad esempio, il valore predefinito per il daemon multipath su RHEL5,5 potrebbe essere il seguente:

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

Questo può essere corretto con i seguenti comandi:

```
[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
```

Mirroring ASM

Il mirroring ASM potrebbe richiedere modifiche alle impostazioni di multipath Linux per consentire ad ASM di riconoscere un problema e passare a un gruppo di errori alternativo. La maggior parte delle configurazioni ASM su ONTAP utilizza la ridondanza esterna, il che significa che la protezione dei dati è fornita dall'array esterno e ASM non esegue il mirroring dei dati. Alcuni siti utilizzano ASM con ridondanza normale per fornire il mirroring bidirezionale, in genere su siti diversi.

Le impostazioni di Linux visualizzate nella "Documentazione delle utilità host NetApp" Includi parametri multipath che determinano indefinite code di i/O. Ciò significa che un i/o su un dispositivo LUN senza percorsi attivi attende finché l'i/o non viene completato. Questo è solitamente consigliabile perché gli host Linux attendono il tempo necessario per il completamento delle modifiche al percorso SAN, per il riavvio degli switch FC o per il completamento di un failover da parte di un sistema di storage.

Questo comportamento di accodamento illimitato causa un problema con il mirroring ASM perché ASM deve ricevere un errore di i/o per consentire al reparto IT di riprovare l'i/o su un LUN alternativo.

Impostare i seguenti parametri in Linux `multipath.conf` File per i LUN ASM utilizzati con il mirroring ASM:

```
polling_interval 5
no_path_retry 24
```

Queste impostazioni creano un timeout di 120 secondi per i dispositivi ASM. Il timeout viene calcolato come `polling_interval` * `no_path_retry` in pochi secondi. In alcuni casi potrebbe essere necessario regolare il valore esatto, ma per la maggior parte degli utilizzi dovrebbe essere sufficiente un timeout di 120 secondi. In particolare, 120 secondi devono consentire il takeover o il giveback del controller senza produrre un errore di i/o che porterebbe il gruppo guasto a diventare offline.

Un più basso `no_path_retry` Il valore può ridurre il tempo richiesto per ASM per passare a un gruppo di errori alternativo, ma aumenta anche il rischio di un failover indesiderato durante attività di manutenzione come il takeover di un controller. Il rischio può essere mitigato tramite un attento monitoraggio dello stato di mirroring

ASM. Se si verifica un failover indesiderato, è possibile risincronizzare rapidamente i mirror se la risincronizzazione viene eseguita in modo relativamente rapido. Per ulteriori informazioni, consultare la documentazione Oracle su ASM Fast Mirror Resync per la versione del software Oracle in uso.

Linux xfs, ext3, e ext4 opzioni di mount

NetApp recommended usando le opzioni di mount predefinite.

ASMILib/AFD (driver filtro ASM)

Argomenti di configurazione specifici per il sistema operativo Linux utilizzando AFD e ASMlib

Dimensioni dei blocchi ASMlib

ASMlib è una libreria di gestione ASM opzionale e le utilità associate. Il suo valore principale è la capacità di contrassegnare un LUN o un file basato su NFS come una risorsa ASM con un'etichetta leggibile da un utente.

Le versioni recenti di ASMlib rilevano un parametro LUN chiamato Logical Blocks per Physical Block Exponent (LBPPBE). Questo valore non è stato segnalato dal target SCSI ONTAP fino a poco tempo fa. Ora restituisce un valore che indica che è preferibile una dimensione blocco 4KB. Questa non è una definizione della dimensione del blocco, ma è un suggerimento per qualsiasi applicazione che utilizza LBPPBE che i/o di una certa dimensione potrebbero essere gestiti in modo più efficiente. ASMlib, tuttavia, interpreta LBPPBE come dimensione del blocco e contrassegna in modo permanente l'intestazione ASM quando viene creato il dispositivo ASM.

Questo processo può causare problemi di aggiornamento e migrazione in vari modi, tutti basati sull'impossibilità di combinare dispositivi ASMlib con dimensioni dei blocchi diverse nello stesso gruppo di dischi ASM.

Ad esempio, gli array meno recenti generalmente riportavano un valore LBPPBE pari a 0 o non riportavano affatto questo valore. ASMlib lo interpreta come una dimensione di blocco di 512 byte. Gli array più recenti dovrebbero essere interpretati come aventi una dimensione del blocco di 4KB KB. Non è possibile combinare dispositivi a 512 byte e 4KB nello stesso gruppo di dischi ASM. In questo modo, si impedirebbe a un utente di aumentare le dimensioni del gruppo di dischi ASM utilizzando LUN di due array o sfruttando ASM come strumento di migrazione. In altri casi, RMAN potrebbe non consentire la copia dei file tra un gruppo di dischi ASM con dimensioni del blocco di 512 byte e un gruppo di dischi ASM con dimensioni del blocco di 4KB KB.

La soluzione preferita è quella di tamponare ASMlib. L'ID del bug di Oracle è 13999609 e la patch è presente in oracleasm-support-2,1.8-1 e versioni successive. Questo patch consente all'utente di impostare il parametro `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` a. `true` in `/etc/sysconfig/oracleasm` file di configurazione. In questo modo, ASMlib non utilizza il parametro LBPPBE, il che significa che i LUN del nuovo array sono ora riconosciuti come dispositivi a blocchi da 512 byte.

L'opzione non modifica le dimensioni del blocco sui LUN precedentemente contrassegnati da ASMlib. Ad esempio, se un gruppo di dischi ASM con blocchi da 512 byte deve essere migrato in un nuovo sistema di storage che riporta un blocco da 4KB KB, è possibile scegliere questa opzione `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Deve essere impostato prima che i nuovi LUN siano contrassegnati con ASMlib. Se i dispositivi sono già stati contrassegnati da oracleasm, è necessario riformattarli prima di essere contrassegnati con una nuova dimensione del blocco. Innanzitutto, deconfigurare il dispositivo con `oracleasm deletedisk`, E quindi cancellare i primi 1GB del dispositivo con `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Infine, se il dispositivo era stato precedentemente partizionato, utilizzare `kpartx` Per rimuovere le partizioni obsolete o semplicemente riavviare il sistema operativo.

Se ASMlib non può essere aggiornato, ASMlib può essere rimosso dalla configurazione. Questa modifica comporta un'interruzione e richiede la rimozione dello stampaggio dei dischi ASM e la verifica che `asm_diskstring` parametro impostato correttamente. Questa modifica, tuttavia, non richiede la migrazione dei dati.

Dimensioni blocco comando filtro ASM (AFD)

AFD è una libreria di gestione ASM opzionale che sta diventando il sostituto di ASMlib. Dal punto di vista dello storage, è molto simile ad ASMlib, ma include funzionalità aggiuntive come la capacità di bloccare i/o non Oracle per ridurre le possibilità di errori di utenti o applicazioni che potrebbero danneggiare i dati.

Dimensioni dei blocchi dei dispositivi

Come ASMlib, anche AFD legge il parametro LUN Logical Blocks per Physical Block Exponent (LBPPBE) e per impostazione predefinita utilizza la dimensione fisica del blocco, non la dimensione logica del blocco.

Ciò potrebbe creare un problema se l'AFD viene aggiunto a una configurazione esistente in cui i dispositivi ASM sono già formattati come dispositivi a blocchi da 512 byte. Il driver AFD riconosce il LUN come un dispositivo 4K e la mancata corrispondenza tra l'etichetta ASM e il dispositivo fisico impedirebbe l'accesso. Allo stesso modo, le migrazioni sarebbero influenzate dal fatto che non è possibile combinare dispositivi a 512 byte e 4KB nello stesso gruppo di dischi ASM. In questo modo, si impedirebbe a un utente di aumentare le dimensioni del gruppo di dischi ASM utilizzando LUN di due array o sfruttando ASM come strumento di migrazione. In altri casi, RMAN potrebbe non consentire la copia dei file tra un gruppo di dischi ASM con dimensioni del blocco di 512 byte e un gruppo di dischi ASM con dimensioni del blocco di 4KB KB.

La soluzione è semplice: AFD include un parametro per controllare se utilizza le dimensioni del blocco logico o fisico. Si tratta di un parametro globale che interessa tutti i dispositivi del sistema. Per forzare AFD a utilizzare le dimensioni del blocco logico, impostare `options oracleafd oracleafd_use_logical_block_size=1` in `/etc/modprobe.d/oracleafd.conf` file.

Dimensioni di trasferimento multipath

Le recenti modifiche al kernel linux impongono restrizioni delle dimensioni di i/o inviate ai dispositivi multipath e AFD non rispetta queste restrizioni. Gli i/o vengono quindi rifiutati, il che causa la disconnessione del percorso LUN. Il risultato è un'impossibilità di installare Oracle Grid, configurare ASM o creare un database.

La soluzione consiste nel specificare manualmente la lunghezza massima di trasferimento nel file multipath.conf per i LUN ONTAP:

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

Anche se attualmente non esistono problemi, questo parametro deve essere impostato se si utilizza AFD per garantire che un futuro aggiornamento linux non causi inaspettatamente problemi.

Microsoft Windows

Argomenti di configurazione per database Oracle su Microsoft Windows con ONTAP.

NFS

Oracle supporta l'utilizzo di Microsoft Windows con il client NFS diretto. Questa funzionalità offre un percorso per i vantaggi di gestione di NFS, tra cui la possibilità di visualizzare i file tra più ambienti, ridimensionare dinamicamente i volumi e sfruttare un protocollo IP meno costoso. Consultare la documentazione ufficiale di Oracle per informazioni sull'installazione e la configurazione di un database in Microsoft Windows utilizzando DNFS. Non esistono Best practice speciali.

SAN

Per un'efficienza di compressione ottimale, assicurarsi che il file system NTFS utilizzi un'unità di allocazione di 8K GB o superiore. L'utilizzo di un'unità di allocazione 4K, generalmente predefinita, influisce negativamente sull'efficienza della compressione.

Solaris

Argomenti di configurazione specifici di Solaris.

Opzioni di montaggio NFS Solaris

Nella tabella seguente sono elencate le opzioni di montaggio di Solaris NFS per una singola istanza.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	`rw,bg,hard,[vers=3,vers=4.1], roto=tcp, timeo=600,rsize=262144,wsize=262144`
File di controllo File di dati Registri di ripristino	`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`

L'utilizzo di `llock` è stato dimostrato di migliorare drasticamente le performance negli ambienti dei clienti rimuovendo la latenza associata all'acquisizione e al rilascio di blocchi sul sistema storage. Utilizzare questa opzione con attenzione negli ambienti in cui sono configurati numerosi server per montare gli stessi file system e Oracle è configurato per montare questi database. Sebbene si tratti di una configurazione molto insolita, viene utilizzata da un numero limitato di clienti. Se un'istanza viene avviata accidentalmente una seconda volta, i dati potrebbero danneggiarsi perché Oracle non è in grado di rilevare i file di blocco sul server esterno. I blocchi NFS non offrono altrimenti protezione; come nella versione 3 di NFS, sono solo di natura consultiva.

Perché il `llock` e. `forcedirectio` i parametri si escludono a vicenda, è importante che `filesystemio_options=setall` è presente in `init.ora` file in modo che `directio` viene utilizzato. Senza questo parametro, viene utilizzato il caching del buffer del sistema operativo host e le prestazioni possono essere compromesse.

Nella tabella seguente sono elencate le opzioni di montaggio Solaris NFS RAC.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac`
File di controllo File di dati Registri di ripristino	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio`
CRS/votazione	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio`
Dedicato `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Condiviso `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,suid`

L'aggiunta fa la differenza principale tra le opzioni di montaggio RAC e a istanza singola `noac` e. `forcedirectio` alle opzioni di montaggio. Questa aggiunta ha l'effetto di disabilitare il caching del sistema operativo host, consentendo a tutte le istanze nel cluster RAC di avere una visione coerente dello stato dei dati. Anche se si utilizza il `init.ora` parametro `filesystemio_options=setall` ha lo stesso effetto di disabilitare la cache dell'host, è comunque necessario utilizzare `noac` e. `forcedirectio`.

Il motivo `actimeo=0` è obbligatorio per condiviso `ORACLE_HOME` Le distribuzioni consentono di semplificare la coerenza di file quali file di password Oracle e file spfile. Se ogni istanza di un cluster RAC dispone di un'istanza dedicata `ORACLE_HOME`, questo parametro non è richiesto.

Opzioni di montaggio UFS di Solaris

NetApp consiglia vivamente di utilizzare l'opzione di montaggio della registrazione in modo che l'integrità dei dati venga preservata in caso di arresto anomalo dell'host Solaris o di interruzione della connettività FC. L'opzione di montaggio della registrazione preserva anche l'usabilità dei backup Snapshot.

Solaris ZFS

Solaris ZFS deve essere installato e configurato con attenzione per garantire prestazioni ottimali.

mvector

Solaris 11 ha introdotto una modifica nel modo in cui elabora operazioni i/o di grandi dimensioni, che può causare gravi problemi di prestazioni sugli array di storage SAN. Il problema è documentato nel rapporto 630173 del bug di monitoraggio di NetApp, "regressione delle prestazioni di Solaris 11 ZFS".

Questo non è un bug di ONTAP. Si tratta di un difetto di Solaris rilevato in Solaris Defects 7199305 e 7082975.

È possibile consultare il supporto Oracle per scoprire se la versione di Solaris 11 in uso è interessata o per verificare la soluzione alternativa, passando `zfs_mvector_max_size` a un valore inferiore.

È possibile farlo eseguendo il seguente comando come root:

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

Se da questa modifica emergono problemi imprevisti, è possibile annullarli facilmente eseguendo il seguente comando come root:

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

Kernel

Prestazioni ZFS affidabili richiedono un kernel Solaris con patch contro i problemi di allineamento LUN. La correzione è stata introdotta con la patch 147440-19 in Solaris 10 e con SRU 10,5 per Solaris 11. Utilizzare solo Solaris 10 e versioni successive con ZFS.

Configurazione del LUN

Per configurare un LUN, attenersi alla seguente procedura:

1. Creare un LUN di tipo `solaris`.
2. Installare l'host Utility Kit (HUK) appropriato specificato da "Tool di matrice di interoperabilità NetApp (IMT)".
3. Seguire esattamente le istruzioni nell'HUK come descritto. I passaggi di base sono descritti di seguito, ma fare riferimento a. "documentazione più recente" per la procedura corretta.
   a. Eseguire `host_config` utilità per aggiornare `sd.conf/sdd.conf` file. Questo consente alle unità SCSI di rilevare correttamente i LUN ONTAP.
   b. Seguire le istruzioni fornite da `host_config` Utility per abilitare l'input/output multipath (MPIO).
   c. Reboot (Riavvia). Questa fase è necessaria per consentire il riconoscimento di eventuali modifiche nel sistema.
4. Partizionare i LUN e verificare che siano allineati correttamente. Vedere "Appendice B: Verifica dell'allineamento WAFL" per istruzioni su come eseguire direttamente il test e confermare l'allineamento.

zpool

Uno zpool deve essere creato solo dopo i passaggi nella "Configurazione LUN" vengono eseguite. Se la procedura non viene eseguita correttamente, le prestazioni potrebbero peggiorare notevolmente a causa dell'allineamento i/O. Per ottenere prestazioni ottimali con ONTAP è necessario allineare l'i/o a un confine di 4K su un'unità. I file system creati su uno zpool utilizzano una dimensione di blocco effettiva controllata tramite un parametro chiamato `ashift`, che può essere visualizzato eseguendo il comando `zdb -C`.

Il valore di `ashift` il valore predefinito è 9, ovvero 2^9 o 512 byte. Per prestazioni ottimali, la `ashift` Il valore deve essere 12 (2^12=4K). Questo valore viene impostato al momento della creazione di zpool e non può essere modificato, il che significa che i dati in zpool con `ashift` oltre a 12 deve essere eseguita la migrazione copiando i dati in uno zpool appena creato.

Dopo aver creato uno zpool, verificare il valore di `ashift` prima di procedere. Se il valore non è 12, i LUN non sono stati rilevati correttamente. Distruggere lo zpool, verificare che tutti i passaggi indicati nella relativa documentazione delle utilità host siano stati eseguiti correttamente e ricreare lo zpool.

Zpool e LDOM Solaris

Gli LDOM di Solaris creano un requisito aggiuntivo per assicurarsi che l'allineamento i/o sia corretto. Sebbene un LUN possa essere rilevato correttamente come un dispositivo 4K, un dispositivo vdsk virtuale su un LDOM non eredita la configurazione dal dominio i/O. Vdsk basato su tale LUN torna per impostazione predefinita a un blocco da 512 byte.

È necessario un file di configurazione aggiuntivo. In primo luogo, i singoli LDOM devono essere aggiornati per Oracle bug 15824910 per abilitare le opzioni di configurazione aggiuntive. Questa patch è stata trasferita in tutte le versioni attualmente utilizzate di Solaris. Una volta installato il software LDOM, è pronto per la configurazione dei nuovi LUN correttamente allineati come segue:

1. Identificare il LUN o i LUN da utilizzare nel nuovo zpool. In questo esempio, si tratta del dispositivo c2d1.

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

2. Recuperare l'istanza vdc dei dispositivi da utilizzare per un pool ZFS:

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

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

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

Ciò significa che all'istanza di dispositivo 1 viene assegnata una dimensione di blocco di 4096.

Come ulteriore esempio, si supponga che le istanze vdsk da 1 a 6 debbano essere configurate per una dimensione di blocco di 4K e. `/etc/path_to_inst` recita:

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

4. La finale `vdc.conf` il file deve contenere quanto segue:

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

Attenzione
L'LDOM deve essere riavviato dopo la configurazione di vdc.conf e la creazione di vdsk. Questa fase non può essere evitata. La modifica delle dimensioni del blocco ha effetto solo dopo un riavvio. Procedere con la configurazione di zpool e accertarsi che l'ashift sia impostato correttamente su 12 come descritto in precedenza.

ZFS Intent Log (ZIL)

In genere, non esiste alcun motivo per individuare ZFS Intent Log (ZIL) su un dispositivo diverso. Il registro può condividere lo spazio con il pool principale. L'uso principale di una ZIL separata è quando si utilizzano unità fisiche che non dispongono delle funzionalità di cache di scrittura nei moderni array di storage.

logbias

Impostare `logbias` Parametro sui file system ZFS che ospitano dati Oracle.

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

L'utilizzo di questo parametro riduce i livelli di scrittura complessivi. Per impostazione predefinita, i dati scritti vengono salvati prima nella ZIL e quindi nel pool di storage principale. Questo approccio è appropriato per una configurazione che utilizza una configurazione a disco normale, che include un dispositivo ZIL basato su SSD e supporti rotanti per il pool di storage principale. Questo perché consente l'esecuzione di un commit in una singola transazione i/o sul supporto con latenza più bassa disponibile.

Quando si utilizza un moderno storage array che include funzionalità di caching autonome, questo approccio generalmente non è necessario. In rare circostanze, potrebbe essere opportuno assegnare una scrittura con una singola transazione al registro, ad esempio un carico di lavoro costituito da scritture casuali altamente concentrate e sensibili alla latenza. Vi sono conseguenze sotto forma di amplificazione in scrittura poiché i dati registrati vengono infine scritti nel pool di archiviazione principale, con il risultato di raddoppiare l'attività di scrittura.

I/o diretto

Molte applicazioni, inclusi i prodotti Oracle, possono bypassare la cache del buffer host attivando l'i/o diretto Questa strategia non funziona come previsto con i file system ZFS. Anche se la cache del buffer host viene ignorata, ZFS continua a memorizzare i dati nella cache. Questa azione può produrre risultati fuorvianti quando si utilizzano strumenti come fio o sio per eseguire test delle prestazioni perché è difficile prevedere se l'i/o raggiunge il sistema di storage o se viene memorizzato nella cache locale del sistema operativo. Questa azione rende inoltre molto difficile l'utilizzo di tali test sintetici per confrontare le prestazioni di ZFS con altri file system. In pratica, le performance del file system differiscono da poco a nulla per i carichi di lavoro degli utenti reali.

Diversi zpool

Backup basati su snapshot, ripristini, cloni e archiviazione dei dati basati su ZFS devono essere eseguiti al livello di zpool e in genere richiedono più zpool. Uno zpool è analogo a un gruppo di dischi LVM e deve essere configurato utilizzando le stesse regole. Ad esempio, è probabilmente meglio disporre un database con i file di dati residenti su `zpool1` e i log di archivio, i file di controllo e i log di ripristino che risiedono su `zpool2`. Questo approccio consente un backup a caldo standard in cui il database viene posto in modalità hot backup, seguito da uno snapshot di `zpool1`. Il database viene quindi rimosso dalla modalità di backup a caldo, l'archivio di log viene forzato e viene creata una snapshot di `zpool2` viene creato. Un'operazione di ripristino

richiede lo smontaggio dei file system zfs e l'offlining completo di zpool, in seguito a un'operazione di ripristino di SnapRestore. Lo zpool può quindi essere portato nuovamente online e il database recuperato.

filesystemio_options

Parametro Oracle `filesystemio_options` Funziona in modo diverso con ZFS. Se `setall` oppure `directio` Viene utilizzato, le operazioni di scrittura sono sincrone e ignorano la cache del buffer del sistema operativo, ma le letture sono bufferizzate da ZFS. Questa azione causa difficoltà nell'analisi delle performance perché talvolta l'i/o viene intercettato e gestito dalla cache ZFS, rendendo la latenza dello storage e l'i/o totale inferiori a quanto pare.

Configurazione host con sistemi ASA r2

AIX

Argomenti di configurazione per il database Oracle su IBM AIX con ASA r2 ONTAP.

AIX è supportato con NetApp ASA r2 per l'hosting di database Oracle, a condizione che:

- Configurare Oracle correttamente per l'I/O simultaneo.
- Si utilizzano protocolli SAN supportati (FC/iSCSI/NVMe).
- Si esegue ONTAP 9.16.x o versione successiva su ASA r2.

I/o simultanei

Per ottenere prestazioni ottimali su IBM AIX con ASA r2 è necessario utilizzare I/O simultanei. Senza I/O simultaneo, è probabile che si verifichino limitazioni nelle prestazioni perché AIX esegue I/O atomici serializzati, il che comporta un sovraccarico significativo.

Originariamente, NetApp consigliava di utilizzare `cio` opzione di montaggio per forzare l'I/O simultaneo sul file system, ma questo processo presentava degli svantaggi e non è più necessario. Dall'introduzione di AIX 5.2 e Oracle 10gR1, Oracle su AIX può aprire singoli file per l'I/O simultaneo, anziché forzare l'I/O simultaneo sull'intero file system.

Il metodo migliore per abilitare l'i/o simultaneo è impostare `init.ora` parametro `filesystemio_options` a. `setall`. In questo modo, Oracle può aprire file specifici da utilizzare con i/o simultanei

L'utilizzo di cio come opzione di montaggio forza l'uso di I/O simultanei, il che può avere conseguenze negative. Ad esempio, forzare l'I/O simultaneo disabilita la lettura anticipata sui file system, il che può compromettere le prestazioni dell'I/O che si verifica all'esterno del software del database Oracle, come la copia di file e l'esecuzione di backup su nastro. Inoltre, prodotti come Oracle GoldenGate e SAP BR*Tools non sono compatibili con l'utilizzo dell'opzione di montaggio cio con alcune versioni di Oracle.

NetApp consiglia quanto segue:

- Non utilizzare `cio` opzione di montaggio a livello di file system. Abilitare invece l'i/o simultaneo tramite l'utilizzo di `filesystemio_options=setall`.
- Utilizzare solo il `cio` opzione di montaggio se non è possibile impostare `filesystemio_options=setall`.

Poiché ASA r2 non supporta NAS, tutte le distribuzioni Oracle su AIX devono utilizzare protocolli a blocchi.

Opzioni di montaggio di AIX jfs/JFS2

Nella tabella seguente sono elencate le opzioni di montaggio di AIX jfs/JFS2.

Tipo di file	Opzioni di montaggio
Pagina iniziale ADR	Valori predefiniti
File di controllo	Valori predefiniti
File di dati	Valori predefiniti
Redo log	Valori predefiniti
ORACLE_HOME	Valori predefiniti

Prima di utilizzare AIX `hdisk` dispositivi in qualsiasi ambiente, inclusi i database, controllano il parametro `queue_depth`. Questo parametro non è la profondità della coda HBA; piuttosto è correlato alla profondità della coda SCSI del singolo `hdisk device`. A seconda di come sono configurati i LUN ASA r2, il valore per `queue_depth` potrebbe essere troppo basso per ottenere buone prestazioni. I test hanno dimostrato che il valore ottimale è 64.

HP-UX

Argomenti di configurazione per il database Oracle su HP-UX con ASA r2 ONTAP.

HP-UX è supportato con NetApp ASA r2 per l'hosting di database Oracle, a condizione che:

- La versione ONTAP è 9.16.x o successiva.
- Utilizzare protocolli SAN (FC/iSCSI/NVMe). NAS non è supportato su ASA r2.
- Applicare le best practice di montaggio e ottimizzazione I/O specifiche di HP-UX.

Opzioni di montaggio VxFS HP-UX

Utilizzare le seguenti opzioni di montaggio per i file system che ospitano file binari Oracle:

```
delaylog,nodatainlog
```

Utilizzare le seguenti opzioni di montaggio per i file system contenenti file di dati, log di ripristino, log di archivio e file di controllo in cui la versione di HP-UX non supporta i/o simultanei:

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

Quando l'i/o simultaneo è supportato (VxFS 5.0.1 e versioni successive o con ServiceGuard Storage Management Suite), utilizzare queste opzioni di montaggio per i file system contenenti file di dati, log di ripristino, log di archivio e file di controllo:

```
delaylog,cio
```

> Il parametro `db_file_multiblock_read_count` È particolarmente critico negli ambienti VxFS. Oracle consiglia di non impostare questo parametro in Oracle 10g R1 e versioni successive, a meno che non sia diversamente specificato. L'impostazione predefinita con dimensioni blocco Oracle 8KB è 128 KB. Se il valore di questo parametro è forzato a 16 o inferiore, rimuovere l' `convosync=direct` Montare l'opzione perché può danneggiare le prestazioni i/o sequenziali. Questa operazione danneggia altri aspetti delle prestazioni e deve essere eseguita solo se il valore di `db_file_multiblock_read_count` deve essere modificato dal valore predefinito.

Linux

Argomenti di configurazione specifici del sistema operativo Linux con ASA r2 ONTAP.

> Linux (Oracle Linux, RHEL, SUSE) è supportato con ASA r2 per i database Oracle. Utilizzare protocolli SAN, configurare correttamente il multipathing e applicare le best practice di Oracle per l'ottimizzazione di ASM e I/O.

Utilità di pianificazione i/O.

Il kernel Linux permette un controllo di basso livello sul modo in cui l'i/o blocca i dispositivi è programmato. Le impostazioni predefinite su varie distribuzioni di Linux variano notevolmente. I test dimostrano che la scadenza di solito offre i migliori risultati, ma a volte NOOP è stato leggermente migliore. La differenza di prestazioni è minima, ma è necessario verificare entrambe le opzioni se è necessario estrarre le massime prestazioni possibili da una configurazione di database. CFQ è l'impostazione predefinita in molte configurazioni e ha dimostrato di avere problemi significativi di prestazioni con i carichi di lavoro del database.

Per istruzioni sulla configurazione dello scheduler i/o, consultare la documentazione del fornitore di Linux pertinente.

Multipathing

Alcuni clienti hanno riscontrato arresti anomali durante l'interruzione della rete perché il daemon multipath non era in esecuzione sul proprio sistema. Nelle versioni recenti di Linux, il processo di installazione del sistema operativo e del demone multipathing potrebbero lasciare questi sistemi operativi vulnerabili a questo problema. I pacchetti sono installati correttamente, ma non sono configurati per l'avvio automatico dopo un riavvio.

Ad esempio, l'impostazione predefinita per il demone multipath su RHEL 9.7 potrebbe apparire come segue:

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

Questo può essere corretto con i seguenti comandi:

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

Profondità della coda

Impostare la profondità della coda appropriata per i dispositivi SAN per evitare colli di bottiglia I/O. La profondità della coda predefinita su Linux è spesso impostata su 128, il che può causare problemi di prestazioni con i database Oracle. Impostando una profondità della coda troppo alta si può causare un'eccessiva coda di I/O, con conseguente aumento della latenza e riduzione della produttività. Un valore troppo basso può limitare il numero di richieste I/O in sospeso, riducendo le prestazioni complessive. Una profondità della coda di 64 è spesso un buon punto di partenza per i carichi di lavoro del database Oracle su ASA r2, ma potrebbe essere necessario regolarla in base alle caratteristiche specifiche del carico di lavoro e ai test delle prestazioni.

Mirroring ASM

Il mirroring ASM potrebbe richiedere modifiche alle impostazioni di multipath Linux per consentire ad ASM di riconoscere un problema e passare a un gruppo di errori alternativo. La maggior parte delle configurazioni ASM su ONTAP utilizza la ridondanza esterna, il che significa che la protezione dei dati è fornita dall'array esterno e ASM non esegue il mirroring dei dati. Alcuni siti utilizzano ASM con ridondanza normale per fornire il mirroring bidirezionale, in genere su siti diversi.

Per i sistemi ASA r2 che supportano il multipathing attivo-attivo, è necessario regolare queste impostazioni multipath. Poiché tutti i percorsi sono attivi e con carico bilanciato, non è necessaria una coda indefinita. I parametri multipath dovrebbero invece dare priorità alle prestazioni e al failback rapido. Questo comportamento è importante per il mirroring ASM perché ASM deve ricevere un errore di I/O per poter riprovare l'I/O su una LUN alternativa. Se l'I/O viene messo in coda indefinitamente, ASM non può attivare un failover.

Impostare i seguenti parametri in Linux `multipath.conf` File per i LUN ASM utilizzati con il mirroring ASM:

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

Queste impostazioni creano un timeout di 120 secondi per i dispositivi ASM. Il timeout viene calcolato come `polling_interval` * `no_path_retry` in pochi secondi. In alcuni casi potrebbe essere necessario regolare il valore esatto, ma per la maggior parte degli utilizzi dovrebbe essere sufficiente un timeout di 120 secondi. In particolare, 120 secondi devono consentire il takeover o il giveback del controller senza produrre un errore di i/o che porterebbe il gruppo guasto a diventare offline.

Un più basso `no_path_retry` Il valore può ridurre il tempo richiesto per ASM per passare a un gruppo di errori alternativo, ma aumenta anche il rischio di un failover indesiderato durante attività di manutenzione come il takeover di un controller. Il rischio può essere mitigato tramite un attento monitoraggio dello stato di mirroring ASM. Se si verifica un failover indesiderato, è possibile risincronizzare rapidamente i mirror se la risincronizzazione viene eseguita in modo relativamente rapido. Per ulteriori informazioni, consultare la documentazione Oracle su ASM Fast Mirror Resync per la versione del software Oracle in uso.

Linux xfs, ext3, e ext4 opzioni di mount

* NetApp consiglia* di utilizzare le opzioni di montaggio predefinite. Assicurare il corretto allineamento durante la creazione di file system su LUN.

ASMILib/AFD (driver filtro ASM)

Argomenti di configurazione specifici del sistema operativo Linux mediante AFD e ASMlib con ASA r2 ONTAP.

Dimensioni dei blocchi ASMlib

ASMlib è una libreria di gestione ASM opzionale e relative utilità. Il suo valore principale è la capacità di contrassegnare una LUN come risorsa ASM con un'etichetta leggibile dall'uomo.

Le versioni recenti di ASMlib rilevano un parametro LUN chiamato Logical Blocks per Physical Block Exponent (LBPPBE). Questo valore non è stato segnalato dal target SCSI ONTAP fino a poco tempo fa. Ora restituisce un valore che indica che è preferibile una dimensione blocco 4KB. Questa non è una definizione della dimensione del blocco, ma è un suggerimento per qualsiasi applicazione che utilizza LBPPBE che i/o di una certa dimensione potrebbero essere gestiti in modo più efficiente. ASMlib, tuttavia, interpreta LBPPBE come dimensione del blocco e contrassegna in modo permanente l'intestazione ASM quando viene creato il dispositivo ASM.

Questo processo può causare problemi di aggiornamento e migrazione in vari modi, tutti basati sull'impossibilità di combinare dispositivi ASMlib con dimensioni dei blocchi diverse nello stesso gruppo di dischi ASM.

Ad esempio, gli array meno recenti generalmente riportavano un valore LBPPBE pari a 0 o non riportavano affatto questo valore. ASMlib lo interpreta come una dimensione di blocco di 512 byte. Gli array più recenti dovrebbero essere interpretati come aventi una dimensione del blocco di 4KB KB. Non è possibile combinare dispositivi a 512 byte e 4KB nello stesso gruppo di dischi ASM. In questo modo, si impedirebbe a un utente di aumentare le dimensioni del gruppo di dischi ASM utilizzando LUN di due array o sfruttando ASM come strumento di migrazione. In altri casi, RMAN potrebbe non consentire la copia dei file tra un gruppo di dischi ASM con dimensioni del blocco di 512 byte e un gruppo di dischi ASM con dimensioni del blocco di 4KB KB.

La soluzione preferita è quella di tamponare ASMlib. L'ID del bug di Oracle è 13999609 e la patch è presente in oracleasm-support-2,1.8-1 e versioni successive. Questo patch consente all'utente di impostare il parametro `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` a. `true` in `/etc/sysconfig/oracleasm` file di configurazione. In questo modo, ASMlib non utilizza il parametro LBPPBE, il che significa che i LUN del nuovo array sono ora riconosciuti come dispositivi a blocchi da 512 byte.

L'opzione non modifica le dimensioni del blocco sui LUN precedentemente contrassegnati da ASMlib. Ad esempio, se un gruppo di dischi ASM con blocchi da 512 byte deve essere migrato in un nuovo sistema di storage che riporta un blocco da 4KB KB, è possibile scegliere questa opzione `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Deve essere impostato prima che i nuovi LUN siano contrassegnati con ASMlib. Se i dispositivi sono già stati contrassegnati da oracleasm, è necessario riformattarli prima di essere contrassegnati con una nuova dimensione del blocco. Innanzitutto, deconfigurare il dispositivo con `oracleasm deletedisk`, E quindi cancellare i primi 1GB del dispositivo con `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Infine, se il dispositivo era stato precedentemente partizionato, utilizzare `kpartx` Per rimuovere le partizioni obsolete o semplicemente riavviare il sistema operativo.

Se ASMlib non può essere aggiornato, ASMlib può essere rimosso dalla configurazione. Questa modifica comporta un'interruzione e richiede la rimozione dello stampaggio dei dischi ASM e la verifica che `asm_diskstring` parametro impostato correttamente. Questa modifica, tuttavia, non richiede la migrazione dei dati.

Dimensioni blocco comando filtro ASM (AFD)

AFD è una libreria di gestione ASM opzionale che sta diventando il sostituto di ASMlib. Dal punto di vista dello storage, è molto simile ad ASMlib, ma include funzionalità aggiuntive come la capacità di bloccare i/o non Oracle per ridurre le possibilità di errori di utenti o applicazioni che potrebbero danneggiare i dati.

Dimensioni dei blocchi dei dispositivi

Come ASMlib, anche AFD legge il parametro LUN Logical Blocks per Physical Block Exponent (LBPPBE) e per impostazione predefinita utilizza la dimensione fisica del blocco, non la dimensione logica del blocco.

Ciò potrebbe creare un problema se l'AFD viene aggiunto a una configurazione esistente in cui i dispositivi ASM sono già formattati come dispositivi a blocchi da 512 byte. Il driver AFD riconosce il LUN come un dispositivo 4K e la mancata corrispondenza tra l'etichetta ASM e il dispositivo fisico impedirebbe l'accesso. Allo stesso modo, le migrazioni sarebbero influenzate dal fatto che non è possibile combinare dispositivi a 512 byte e 4KB nello stesso gruppo di dischi ASM. In questo modo, si impedirebbe a un utente di aumentare le dimensioni del gruppo di dischi ASM utilizzando LUN di due array o sfruttando ASM come strumento di migrazione. In altri casi, RMAN potrebbe non consentire la copia dei file tra un gruppo di dischi ASM con dimensioni del blocco di 512 byte e un gruppo di dischi ASM con dimensioni del blocco di 4KB KB.

La soluzione è semplice: AFD include un parametro per controllare se utilizza le dimensioni del blocco logico o fisico. Si tratta di un parametro globale che interessa tutti i dispositivi del sistema. Per forzare AFD a utilizzare le dimensioni del blocco logico, impostare `options oracleafd oracleafd_use_logical_block_size=1` in `/etc/modprobe.d/oracleafd.conf` file.

Dimensioni di trasferimento multipath

Le recenti modifiche al kernel linux impongono restrizioni delle dimensioni di i/o inviate ai dispositivi multipath e AFD non rispetta queste restrizioni. Gli i/o vengono quindi rifiutati, il che causa la disconnessione del percorso LUN. Il risultato è un'impossibilità di installare Oracle Grid, configurare ASM o creare un database.

La soluzione consiste nel specificare manualmente la lunghezza massima di trasferimento nel file multipath.conf per i LUN ONTAP:

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

Anche se attualmente non esistono problemi, questo parametro deve essere impostato se si utilizza AFD per garantire che un futuro aggiornamento linux non causi inaspettatamente problemi.

Microsoft Windows

Argomenti di configurazione per il database Oracle su Microsoft Windows con ASA r2 ONTAP.

SAN

Per un'efficienza di compressione ottimale, assicurarsi che il file system NTFS utilizzi un'unità di allocazione di 8K GB o superiore. L'utilizzo di un'unità di allocazione 4K, generalmente predefinita, influisce negativamente sull'efficienza della compressione.

Solaris

Argomenti di configurazione specifici del sistema operativo Solaris con ASA r2 ONTAP.

Opzioni di montaggio UFS di Solaris

NetApp consiglia vivamente di utilizzare l'opzione di montaggio della registrazione in modo che l'integrità dei dati venga preservata in caso di arresto anomalo dell'host Solaris o di interruzione della connettività FC. L'opzione di montaggio della registrazione preserva anche l'usabilità dei backup Snapshot.

Solaris ZFS

Solaris ZFS deve essere installato e configurato con attenzione per garantire prestazioni ottimali.

mvector

Solaris 11 ha introdotto una modifica nel modo in cui elabora operazioni i/o di grandi dimensioni, che può causare gravi problemi di prestazioni sugli array di storage SAN. Il problema è documentato nel rapporto 630173 del bug di monitoraggio di NetApp, "regressione delle prestazioni di Solaris 11 ZFS".

Questo non è un bug di ONTAP. Si tratta di un difetto di Solaris rilevato in Solaris Defects 7199305 e 7082975.

È possibile consultare il supporto Oracle per scoprire se la versione di Solaris 11 in uso è interessata o per verificare la soluzione alternativa, passando `zfs_mvector_max_size` a un valore inferiore.

È possibile farlo eseguendo il seguente comando come root:

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

Se da questa modifica emergono problemi imprevisti, è possibile annullarli facilmente eseguendo il seguente comando come root:

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

Kernel

Prestazioni ZFS affidabili richiedono un kernel Solaris con patch contro i problemi di allineamento LUN. La correzione è stata introdotta con la patch 147440-19 in Solaris 10 e con SRU 10,5 per Solaris 11. Utilizzare solo Solaris 10 e versioni successive con ZFS.

Configurazione del LUN

Per configurare un LUN, attenersi alla seguente procedura:

1. Creare un LUN di tipo `solaris`.
2. Installare l'host Utility Kit (HUK) appropriato specificato da "Tool di matrice di interoperabilità NetApp (IMT)".
3. Seguire esattamente le istruzioni nell'HUK come descritto. I passaggi di base sono descritti di seguito, ma fare riferimento a. "documentazione più recente" per la procedura corretta.
    a. Eseguire `host_config` utilità per aggiornare `sd.conf/sdd.conf` file. Questo consente alle unità SCSI di rilevare correttamente i LUN ONTAP.
    b. Seguire le istruzioni fornite da `host_config` Utility per abilitare l'input/output multipath (MPIO).
    c. Reboot (Riavvia). Questa fase è necessaria per consentire il riconoscimento di eventuali modifiche nel sistema.
4. Partizionare i LUN e verificare che siano allineati correttamente. Vedere "Appendice B: Verifica dell'allineamento WAFL" per istruzioni su come eseguire direttamente il test e confermare l'allineamento.

zpool

Uno zpool deve essere creato solo dopo i passaggi nella "Configurazione LUN" vengono eseguite. Se la procedura non viene eseguita correttamente, le prestazioni potrebbero peggiorare notevolmente a causa dell'allineamento i/O. Per ottenere prestazioni ottimali con ONTAP è necessario allineare l'i/o a un confine di 4K su un'unità. I file system creati su uno zpool utilizzano una dimensione di blocco effettiva controllata tramite un parametro chiamato `ashift`, che può essere visualizzato eseguendo il comando `zdb -C`.

Il valore di `ashift` il valore predefinito è 9, ovvero 2^9 o 512 byte. Per prestazioni ottimali, la `ashift` Il valore deve essere 12 (2^12=4K). Questo valore viene impostato al momento della creazione di zpool e non può essere modificato, il che significa che i dati in zpool con `ashift` oltre a 12 deve essere eseguita la migrazione copiando i dati in uno zpool appena creato.

Dopo aver creato uno zpool, verificare il valore di `ashift` prima di procedere. Se il valore non è 12, i LUN non sono stati rilevati correttamente. Distruggere lo zpool, verificare che tutti i passaggi indicati nella relativa documentazione delle utilità host siano stati eseguiti correttamente e ricreare lo zpool.

Zpool e LDOM Solaris

Gli LDOM di Solaris creano un requisito aggiuntivo per assicurarsi che l'allineamento i/o sia corretto. Sebbene un LUN possa essere rilevato correttamente come un dispositivo 4K, un dispositivo vdsk virtuale su un LDOM non eredita la configurazione dal dominio i/O. Vdsk basato su tale LUN torna per impostazione predefinita a un blocco da 512 byte.

È necessario un file di configurazione aggiuntivo. In primo luogo, i singoli LDOM devono essere aggiornati per Oracle bug 15824910 per abilitare le opzioni di configurazione aggiuntive. Questa patch è stata trasferita in tutte le versioni attualmente utilizzate di Solaris. Una volta installato il software LDOM, è pronto per la configurazione dei nuovi LUN correttamente allineati come segue:

1. Identificare il LUN o i LUN da utilizzare nel nuovo zpool. In questo esempio, si tratta del dispositivo c2d1.

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

2. Recuperare l'istanza vdc dei dispositivi da utilizzare per un pool ZFS:

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

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

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

Ciò significa che all'istanza di dispositivo 1 viene assegnata una dimensione di blocco di 4096.

Come ulteriore esempio, si supponga che le istanze vdsk da 1 a 6 debbano essere configurate per una dimensione di blocco di 4K e. `/etc/path_to_inst` recita:

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

4. La finale `vdc.conf` il file deve contenere quanto segue:

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

L'LDOM deve essere riavviato dopo la configurazione di vdc.conf e la creazione di vdsk. Questa fase non può essere evitata. La modifica delle dimensioni del blocco ha effetto solo dopo un riavvio. Procedere con la configurazione di zpool e accertarsi che l'ashift sia impostato correttamente su 12 come descritto in precedenza.

ZFS Intent Log (ZIL)

In genere, non esiste alcun motivo per individuare ZFS Intent Log (ZIL) su un dispositivo diverso. Il registro può condividere lo spazio con il pool principale. L'uso principale di una ZIL separata è quando si utilizzano unità fisiche che non dispongono delle funzionalità di cache di scrittura nei moderni array di storage.

logbias

Impostare `logbias` Parametro sui file system ZFS che ospitano dati Oracle.

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

L'utilizzo di questo parametro riduce i livelli di scrittura complessivi. Per impostazione predefinita, i dati scritti vengono salvati prima nella ZIL e quindi nel pool di storage principale. Questo approccio è appropriato per una configurazione che utilizza una configurazione a disco normale, che include un dispositivo ZIL basato su SSD e supporti rotanti per il pool di storage principale. Questo perché consente l'esecuzione di un commit in una singola transazione i/o sul supporto con latenza più bassa disponibile.

Quando si utilizza un moderno storage array che include funzionalità di caching autonome, questo approccio generalmente non è necessario. In rare circostanze, potrebbe essere opportuno assegnare una scrittura con una singola transazione al registro, ad esempio un carico di lavoro costituito da scritture casuali altamente concentrate e sensibili alla latenza. Vi sono conseguenze sotto forma di amplificazione in scrittura poiché i dati registrati vengono infine scritti nel pool di archiviazione principale, con il risultato di raddoppiare l'attività di scrittura.

I/o diretto

Molte applicazioni, inclusi i prodotti Oracle, possono bypassare la cache del buffer host attivando l'i/o diretto Questa strategia non funziona come previsto con i file system ZFS. Anche se la cache del buffer host viene ignorata, ZFS continua a memorizzare i dati nella cache. Questa azione può produrre risultati fuorvianti quando si utilizzano strumenti come fio o sio per eseguire test delle prestazioni perché è difficile prevedere se

l'i/o raggiunge il sistema di storage o se viene memorizzato nella cache locale del sistema operativo. Questa azione rende inoltre molto difficile l'utilizzo di tali test sintetici per confrontare le prestazioni di ZFS con altri file system. In pratica, le performance del file system differiscono da poco a nulla per i carichi di lavoro degli utenti reali.

Diversi zpool

Backup basati su snapshot, ripristini, cloni e archiviazione dei dati basati su ZFS devono essere eseguiti al livello di zpool e in genere richiedono più zpool. Uno zpool è analogo a un gruppo di dischi LVM e deve essere configurato utilizzando le stesse regole. Ad esempio, è probabilmente meglio disporre un database con i file di dati residenti su `zpool1` e i log di archivio, i file di controllo e i log di ripristino che risiedono su `zpool2`. Questo approccio consente un backup a caldo standard in cui il database viene posto in modalità hot backup, seguito da uno snapshot di `zpool1`. Il database viene quindi rimosso dalla modalità di backup a caldo, l'archivio di log viene forzato e viene creata una snapshot di `zpool2` viene creato. Un'operazione di ripristino richiede lo smontaggio dei file system zfs e l'offlining completo di zpool, in seguito a un'operazione di ripristino di SnapRestore. Lo zpool può quindi essere portato nuovamente online e il database recuperato.

filesystemio_options

Parametro Oracle `filesystemio_options` Funziona in modo diverso con ZFS. Se `setall` oppure `directio` Viene utilizzato, le operazioni di scrittura sono sincrone e ignorano la cache del buffer del sistema operativo, ma le letture sono bufferizzate da ZFS. Questa azione causa difficoltà nell'analisi delle performance perché talvolta l'i/o viene intercettato e gestito dalla cache ZFS, rendendo la latenza dello storage e l'i/o totale inferiori a quanto pare.

Configurazione di rete su sistemi AFF/ FAS

Interfacce logiche

I database Oracle devono accedere allo storage. Le interfacce logiche (LIF) sono le tubazioni di rete che collegano una Storage Virtual Machine (SVM) alla rete e a loro volta al database. La corretta progettazione della LIF è necessaria per garantire una larghezza di banda sufficiente per ogni carico di lavoro del database e il failover non comporta una perdita dei servizi storage.

Questa sezione offre una panoramica dei principali principi di progettazione della LIF. Per una documentazione più completa, vedere "Documentazione di gestione della rete ONTAP". Come per altri aspetti dell'architettura dei database, le migliori opzioni per la progettazione di una Storage Virtual Machine (SVM, nota come vserver all'interfaccia della CLI) e di un'interfaccia logica (LIF) dipendono in gran parte dai requisiti di scalabilità e dalle esigenze di business.

Durante la creazione di una strategia LIF, prendi in considerazione i seguenti argomenti principali:

- **Performance.** la larghezza di banda della rete è sufficiente?
- **Resilienza.** ci sono singoli punti di guasto nel progetto?
- **Gestibilità.** la rete può essere scalata senza interruzioni?

Gli argomenti trattati sono relativi alla soluzione end-to-end, dall'host fino agli switch fino al sistema storage.

Tipi di LIF

Esistono diversi tipi di LIF. "Documentazione ONTAP sui tipi di LIF" Fornisci informazioni più complete su questo argomento, ma da un punto di vista funzionale le LIF possono essere divise in gruppi:

- **LIF di gestione cluster e nodi.** LIF utilizzati per gestire il cluster storage.
- **LIF di gestione SVM.** interfacce che consentono l'accesso a una SVM tramite l'API REST o ONTAPI (nota anche come ZAPI) per funzioni come la creazione di snapshot o il ridimensionamento del volume. Prodotti come SnapManager for Oracle (SMO) devono avere accesso a una LIF di gestione SVM.
- **Interfacce LIF dati** per FC, iSCSI, NVMe/FC, NVMe/TCP, NFS, o dati SMB/CIFS.

> Una LIF dati utilizzata per il traffico NFS può anche essere utilizzata per la gestione cambiando la policy del firewall da `data` a. `mgmt` O un'altra policy che consente HTTP, HTTPS o SSH. Questa modifica può semplificare la configurazione di rete evitando la configurazione di ciascun host per l'accesso sia alla LIF dati NFS che a una LIF di gestione separata. Non è possibile configurare un'interfaccia sia per iSCSI che per il traffico di gestione, nonostante entrambi utilizzino un protocollo IP. Negli ambienti iSCSI è necessaria una LIF di gestione separata.

Progettazione della SAN LIF

Il design di LIF in un ambiente SAN è relativamente semplice per un motivo: Il multipathing. Tutte le moderne implementazioni SAN consentono a un client di accedere ai dati su più percorsi di rete indipendenti e di selezionare i percorsi migliori per l'accesso. Di conseguenza, le performance rispetto alla progettazione LIF sono più semplici da gestire, perché i client SAN bilanciano automaticamente il carico dell'i/o nei migliori percorsi disponibili.

Se un percorso non è disponibile, il client seleziona automaticamente un percorso diverso. Grazie alla sua semplicità di progettazione, le LIF SAN sono generalmente più gestibili. Ciò non significa che un ambiente SAN sia sempre più facile da gestire, poiché vi sono molti altri aspetti dello storage SAN che sono molto più complicati di NFS. Significa semplicemente che la progettazione della SAN LIF è più semplice.

Performance

La considerazione più importante riguardo le performance di una LIF in un ambiente SAN è la larghezza di banda. Ad esempio, un cluster ONTAP AFF a due nodi con due porte FC da 16GB GB per nodo offre fino a 32GB Gbps di larghezza di banda da/per ciascun nodo.

Resilienza

Le LIF SAN non eseguono il failover su un sistema storage AFF. In caso di guasto di una LIF SAN a causa del failover del controller, il software multipath del client rileva la perdita di un percorso e reindirizza l'i/o a una diversa LIF. Con i sistemi storage ASA, il failover delle LIF dopo un breve ritardo, ma ciò non interrompe l'io perché ci sono percorsi già attivi sull'altro controller. Il processo di failover viene eseguito per ripristinare l'accesso dell'host su tutte le porte definite.

Gestibilità

La migrazione LIF è un task molto più comune in un ambiente NFS, perché la migrazione LIF è spesso associata alla riallocazione dei volumi nel cluster. Non è necessario migrare una LIF in un ambiente SAN quando i volumi vengono ricollocati nella coppia ha. Questo perché, una volta completato lo spostamento del volume, ONTAP invia una notifica alla SAN in merito a una modifica dei percorsi e i client SAN vengono automaticamente risottimizzati. La migrazione LIF con SAN è associata principalmente a importanti modifiche hardware fisiche. Ad esempio, per eseguire un upgrade senza interruzioni dei controller, viene eseguita la migrazione di una SAN LIF nel nuovo hardware. Se una porta FC è guasta, una LIF può essere migrata a una

porta inutilizzata.

Raccomandazioni di progettazione

NetApp formula i seguenti consigli:

- Non creare più percorsi di quelli richiesti. Un numero eccessivo di percorsi complica la gestione complessiva e può causare problemi con il failover del percorso su alcuni host. Inoltre, alcuni host hanno limitazioni inattese del percorso per configurazioni come l'avvio SAN.
- Un numero molto ridotto di configurazioni deve richiedere più di quattro percorsi a un LUN. Il valore di avere più di due nodi che pubblicizzano i percorsi delle LUN è limitato perché l'aggregato che ospita un LUN è inaccessibile in caso di guasto del nodo proprietario del LUN e del partner ha. In una situazione del genere, la creazione di percorsi su nodi diversi dalla coppia ha primaria non è d'aiuto.
- Sebbene il numero di percorsi LUN visibili possa essere gestito selezionando le porte incluse nelle zone FC, in genere è più semplice includere tutti i potenziali punti di destinazione nella zona FC e controllare la visibilità delle LUN a livello ONTAP.
- In ONTAP 8,3 e versioni successive, la funzione SLM (Selective LUN mapping) è quella predefinita. Con SLM, ogni nuova LUN viene automaticamente pubblicizzata dal nodo proprietario dell'aggregato sottostante e del partner ha del nodo. Questa disposizione evita la necessità di creare set di porte o configurare la suddivisione in zone per limitare l'accessibilità delle porte. Ogni LUN è disponibile sul numero minimo di nodi necessari per performance e resilienza ottimali.
  *Nel caso in cui sia necessario migrare un LUN all'esterno dei due controller, è possibile aggiungere i nodi aggiuntivi con `lun mapping add-reporting-nodes` In modo che le LUN vengano pubblicizzate sui nuovi nodi. In questo modo si creano ulteriori percorsi SAN alle LUN per la migrazione delle LUN. Tuttavia, l'host deve eseguire un'operazione di rilevamento per utilizzare i nuovi percorsi.
- Non preoccupatevi eccessivamente del traffico indiretto. Si consiglia di evitare il traffico indiretto in un ambiente i/o-intensive per il quale è critico ogni microsecondo di latenza, ma l'effetto visibile delle performance è trascurabile per i workload tipici.

Progettazione della LIF NFS

A differenza dei protocolli SAN, NFS ha una capacità limitata di definire percorsi multipli ai dati. Le estensioni Parallel NFS (pNFS) a NFSv4 risolvono questo limite, ma poiché le velocità ethernet hanno raggiunto 100GB Mbps e oltre, raramente è utile aggiungere altri percorsi.

Performance e resilienza

Sebbene la misurazione delle performance SAN LIF si debba principalmente calcolare la larghezza di banda totale da tutti i percorsi primari, la determinazione delle performance NFS LIF richiede un'analisi più approfondita dell'esatta configurazione di rete. Ad esempio, è possibile configurare due porte 10Gb come porte fisiche grezze oppure come gruppo di interfacce LACP (link Aggregation Control Protocol). Se sono configurati come gruppo di interfacce, sono disponibili più criteri di bilanciamento del carico che funzionano in modo diverso a seconda che il traffico sia commutato o instradato. Infine, Oracle Direct NFS (DNFS) offre configurazioni di bilanciamento del carico attualmente inesistenti in nessun client NFS del sistema operativo.

A differenza dei protocolli SAN, i file system NFS richiedono resilienza al livello del protocollo. Ad esempio, un LUN è sempre configurato con il multipathing attivato, ovvero sono disponibili più canali ridondanti per il sistema storage, ciascuno dei quali utilizza il protocollo FC. Un file system NFS, invece, dipende dalla disponibilità di un unico canale TCP/IP che può essere protetto solo a livello fisico. Questa disposizione è il motivo per cui esistono opzioni quali il failover della porta e l'aggregazione della porta LACP.

In un ambiente NFS, performance e resilienza sono fornite a livello del protocollo di rete. Di conseguenza, entrambi gli argomenti sono intrecciati e devono essere discussi insieme.

Associare le LIF ai gruppi di porte

Per associare una LIF a un gruppo di porte, associare l'indirizzo IP della LIF a un gruppo di porte fisiche. Il metodo principale per aggregare insieme le porte fisiche è LACP. La capacità di fault tolerance di LACP è abbastanza semplice; ogni porta di un gruppo LACP viene monitorata e rimossa dal gruppo di porte in caso di malfunzionamento. Esistono, tuttavia, molte idee sbagliate sul funzionamento di LACP in relazione alle prestazioni:

- LACP non richiede che la configurazione sullo switch corrisponda all'endpoint. Ad esempio, ONTAP può essere configurato con il bilanciamento del carico basato su IP, mentre uno switch può utilizzare il bilanciamento del carico basato su MAC.
- Ogni endpoint che utilizza una connessione LACP può scegliere indipendentemente la porta di trasmissione del pacchetto, ma non può scegliere la porta utilizzata per la ricezione. Ciò significa che il traffico da ONTAP a una destinazione specifica è legato a una porta specifica e il traffico di ritorno potrebbe arrivare su un'interfaccia diversa. Ciò non causa tuttavia problemi.
- LACP non distribuisce uniformemente il traffico in ogni momento. In un ambiente di grandi dimensioni con molti client NFS, il risultato è generalmente l'utilizzo di tutte le porte in un'aggregazione LACP. Tuttavia, qualsiasi file system NFS nell'ambiente è limitato alla larghezza di banda di una sola porta, non all'intera aggregazione.
- Sebbene i criteri LACP di robin-robin siano disponibili su ONTAP, questi criteri non indirizzano la connessione da uno switch a un host. Ad esempio, una configurazione con un trunk LACP a quattro porte su un host e un trunk LACP a quattro porte su ONTAP è ancora in grado di leggere un file system utilizzando una sola porta. Sebbene ONTAP sia in grado di trasmettere dati attraverso tutte e quattro le porte, non sono attualmente disponibili tecnologie di switch che inviano dallo switch all'host attraverso tutte e quattro le porte. Ne viene utilizzato uno solo.

L'approccio più comune in ambienti di grandi dimensioni costituiti da molti host di database è quello di creare un aggregato LACP di un numero appropriato di interfacce 10Gb (o più veloce) utilizzando il bilanciamento del carico IP. Questo approccio consente a ONTAP di garantire l'uso uniforme di tutte le porte, purché esistano un numero sufficiente di client. Il bilanciamento del carico si interrompe quando nella configurazione sono presenti meno client, poiché il trunking LACP non ridistribuisce dinamicamente il carico.

Quando viene stabilita una connessione, il traffico in una determinata direzione viene posizionato su una sola porta. Ad esempio, un database che esegue una scansione completa della tabella su un file system NFS collegato tramite un trunk LACP a quattro porte legge i dati tramite una sola scheda di interfaccia di rete (NIC). Se in un tale ambiente sono presenti solo tre server di database, è possibile che tutti e tre stiano leggendo dalla stessa porta, mentre le altre tre porte sono inattive.

Lega le LIF alle porte fisiche

L'associazione di una LIF a una porta fisica dà come risultato un controllo più granulare della configurazione di rete, in quanto un dato indirizzo IP su un sistema ONTAP è associato a una sola porta di rete alla volta. La resilienza viene quindi ottenuta tramite la configurazione di gruppi di failover e policy di failover.

Criteri di failover e gruppi di failover

Il comportamento delle LIF durante un'interruzione di rete è controllato da policy di failover e gruppi di failover. Le opzioni di configurazione sono state modificate con le diverse versioni di ONTAP. Consultare "Documentazione sulla gestione della rete di ONTAP per gruppi e policy di failover" Per informazioni specifiche sulla versione di ONTAP distribuita.

ONTAP 8,3 (e versioni successive) consente la gestione del failover LIF in base ai domini di broadcast. Pertanto, un amministratore può definire tutte le porte che hanno accesso a una data subnet e consentire a

ONTAP di selezionare una LIF di failover appropriata. Questo approccio può essere utilizzato da alcuni clienti, ma presenta limitazioni in un ambiente di rete di storage ad alta velocità a causa della mancanza di prevedibilità. Ad esempio, un ambiente può includere sia porte 1Gb GbE per l'accesso di routine al file system sia porte 10Gb GbE per l'i/o del file dati Se nello stesso dominio di broadcast sono presenti entrambi i tipi di porte, il failover LIF può spostare l'i/o del file dati da una porta 10Gb a una porta 1Gb.

In sintesi, prendere in considerazione le seguenti pratiche:

1. Configurare un gruppo di failover come definito dall'utente.
2. Popola il gruppo di failover con le porte sul partner controller di failover dello storage (SFO), in modo che le LIF seguano gli aggregati durante un failover dello storage. In questo modo si evita di creare traffico indiretto.
3. Utilizza porte di failover con caratteristiche di performance corrispondenti alla LIF originale. Ad esempio, una LIF su una singola porta fisica di 10Gb deve includere un gruppo di failover con una singola porta 10Gb. Un LIF LACP a quattro porte deve eseguire il failover in un altro LIF LACP a quattro porte. Queste porte sono un sottoinsieme delle porte definite nel dominio di broadcast.
4. Impostare la policy di failover solo su partner SFO. Questo assicura che la LIF segua l'aggregato durante il failover.

Ripristino automatico

Impostare `auto-revert` parametro come desiderato. La maggior parte dei clienti preferisce impostare questo parametro su `true` Di ripristinare la porta home della LIF. Tuttavia, in alcuni casi, i clienti hanno impostato questo valore su `false' per poter esaminare un failover imprevisto prima di restituire una LIF alla porta home.

Rapporto LIF-volume

Un equivoco comune consiste nella necessità di una relazione 1:1:1 tra volumi e LIF NFS. Sebbene questa configurazione sia necessaria per spostare un volume ovunque in un cluster senza creare mai traffico di interconnessione aggiuntivo, non si tratta di un requisito categoricamente importante. Occorre considerare il traffico intercluster, ma la semplice presenza di traffico intercluster non crea problemi. Molti dei benchmark pubblicati per ONTAP includono principalmente l'i/o indiretto

Ad esempio, un progetto di database contenente un numero relativamente contenuto di database critici per le performance, che richiedevano solo un totale di 40 volumi, potrebbe giustificare un volume da 1:1 GB per la strategia LIF, una disposizione che richiederebbe 40 indirizzi IP. Quindi, è possibile spostare un qualsiasi volume nel cluster insieme alla LIF associata e il traffico sarebbe sempre diretto, minimizzando ogni origine di latenza anche a livelli di microsecondi.

Ad esempio, è possibile gestire più facilmente un ambiente di grandi dimensioni in hosting con una relazione di 1:1:1 tra clienti e LIF. Con il passare del tempo, potrebbe essere necessario migrare un volume su un nodo diverso, causando traffico indiretto. Tuttavia, l'effetto sulle prestazioni non dovrebbe essere rilevabile a meno che le porte di rete sullo switch di interconnessione non siano saturanti. In caso di problemi, è possibile stabilire una nuova LIF sui nodi aggiuntivi e l'host può essere aggiornato nella successiva finestra di manutenzione per rimuovere il traffico indiretto dalla configurazione.

Configurazione TCP/IP ed ethernet

Molti clienti di Oracle su ONTAP utilizzano ethernet, il protocollo di rete di NFS, iSCSI, NVMe/TCP e specialmente il cloud.

Impostazioni del sistema operativo host

La maggior parte della documentazione del fornitore di applicazioni include impostazioni TCP ed ethernet specifiche per garantire il funzionamento ottimale dell'applicazione. Queste stesse impostazioni sono in genere sufficienti per fornire anche prestazioni ottimali dello storage basato su IP.

Controllo di flusso Ethernet

Questa tecnologia consente a un client di richiedere che un mittente interrompa temporaneamente la trasmissione dei dati. Questa operazione viene solitamente eseguita perché il ricevitore non è in grado di elaborare i dati in ingresso abbastanza rapidamente. Una volta, la richiesta che un mittente cessi la trasmissione era meno disgregativa di avere pacchetti di scarto del destinatario perché i buffer erano pieni. Questo non è più il caso degli stack TCP utilizzati oggi nei sistemi operativi. Infatti, il controllo di flusso causa più problemi di quanti ne risolva.

Negli ultimi anni sono aumentati i problemi di prestazioni causati dal controllo di flusso Ethernet. Questo perché il controllo di flusso Ethernet opera al livello fisico. Se una configurazione di rete consente a qualsiasi sistema operativo host di inviare una richiesta di controllo di flusso Ethernet a un sistema di storage, il risultato è una pausa in i/o per tutti i client connessi. Poiché un numero crescente di client viene servito da un singolo storage controller, aumenta la probabilità che uno o più client inviino richieste di controllo di flusso. Il problema è stato riscontrato frequentemente presso le sedi dei clienti con un'ampia virtualizzazione del sistema operativo.

Una scheda NIC su un sistema NetApp non dovrebbe ricevere richieste di controllo di flusso. Il metodo utilizzato per ottenere questo risultato varia in base al produttore dello switch di rete. Nella maggior parte dei casi, il controllo di flusso su uno switch Ethernet può essere impostato su `receive desired` oppure `receive on`, il che significa che una richiesta di controllo di flusso non viene inoltrata al controller di memorizzazione. In altri casi, la connessione di rete sul controller di storage potrebbe non consentire la disattivazione del controllo di flusso. In questi casi, i client devono essere configurati in modo da non inviare mai richieste di controllo di flusso, modificando la configurazione NIC sul server host stesso o le porte switch a cui è connesso il server host.

NetApp consiglia assicurarsi che i controller di archiviazione NetApp non ricevano pacchetti di controllo di flusso Ethernet. In genere, è possibile eseguire questa operazione impostando le porte dello switch a cui è collegato il controller, ma alcuni hardware dello switch presentano dei limiti che potrebbero richiedere modifiche sul lato client.

Dimensioni MTU

È stato dimostrato che l'utilizzo dei frame jumbo offre un certo miglioramento delle performance nelle reti 1Gb, riducendo l'overhead della CPU e della rete, ma i benefici non sono solitamente significativi.

NetApp consiglia l'implementazione di frame jumbo quando possibile, sia per ottenere potenziali vantaggi in termini di prestazioni sia per rendere la soluzione a prova di futuro.

L'utilizzo di frame jumbo in una rete 10Gb è quasi obbligatorio. Questo perché la maggior parte delle implementazioni 10Gb raggiungono un limite di pacchetti al secondo senza frame jumbo prima che raggiungano il contrassegno 10Gb. L'utilizzo di frame jumbo migliora l'efficienza dell'elaborazione TCP/IP, poiché consente a sistema operativo, server, schede di rete e sistema di storage di elaborare un numero inferiore di pacchetti, anche se di dimensioni maggiori. Il miglioramento delle prestazioni varia da scheda di rete a scheda di rete, ma è significativo.

Per le implementazioni jumbo-frame, esiste la convinzione comune, ma non corretta, che tutti i dispositivi connessi debbano supportare frame jumbo e che le dimensioni MTU debbano corrispondere end-to-end Al

contrario, i due endpoint di rete negoziano la dimensione del frame più elevata reciprocamente accettabile quando si stabilisce una connessione. In un ambiente tipico, uno switch di rete è impostato su una dimensione MTU di 9216, il controller NetApp è impostato su 9000 e i client sono impostati su una combinazione di 9000 e 1514. I client in grado di supportare un valore MTU di 9000 possono utilizzare frame jumbo, mentre i client in grado di supportare solo 1514 possono negoziare un valore inferiore.

I problemi con questa disposizione sono rari in un ambiente completamente commutato. Tuttavia, in un ambiente con routing occorre assicurarsi che nessun router intermedio sia costretto a frammentare frame jumbo.

NetApp consiglia di configurare quanto segue:

- I frame jumbo sono desiderabili ma non necessari con 1Gb Ethernet (GbE).
- I frame jumbo sono necessari per ottenere le massime prestazioni con 10GbE e velocità.

Parametri TCP

Tre impostazioni spesso non sono configurate correttamente: Timestamp TCP, riconoscimento selettivo (SACK) e ridimensionamento finestra TCP. Molti documenti obsoleti su Internet consigliano di disabilitare uno o più di questi parametri per migliorare le prestazioni. Molti anni fa, questa raccomandazione ha avuto un certo merito quando le capacità della CPU erano molto inferiori e, quando possibile, vi era un vantaggio nel ridurre il sovraccarico sull'elaborazione TCP.

Tuttavia, con i sistemi operativi moderni, la disattivazione di una qualsiasi di queste funzioni TCP in genere non comporta alcun vantaggio rilevabile e, allo stesso tempo, può danneggiare le prestazioni. In ambienti di rete virtualizzati, i danni alle prestazioni sono particolarmente probabili, poiché queste funzioni sono necessarie per gestire in modo efficiente la perdita di pacchetti e le modifiche della qualità della rete.

NetApp consiglia di abilitare timestamp TCP, SACCO e ridimensionamento finestra TCP sull'host, e tutti e tre questi parametri dovrebbero essere attivi per impostazione predefinita in qualsiasi sistema operativo corrente.

Configurazione FC SAN

La configurazione di FC SAN per database Oracle riguarda principalmente le seguenti Best practice quotidiane SAN.

Sono incluse misure di pianificazione tipiche, quali la garanzia della presenza di una larghezza di banda sufficiente sulla SAN tra l'host e il sistema di storage, la verifica della presenza di tutti i percorsi SAN tra i dispositivi richiesti, l'utilizzo delle impostazioni della porta FC richieste dal fornitore dello switch FC, evitando conflitti ISL, e utilizzando un adeguato monitoraggio del fabric SAN.

Suddivisione in zone

Una zona FC non deve mai contenere più di un iniziatore. Una tale disposizione potrebbe sembrare funzionare inizialmente, ma la diafonia tra gli iniziatori interferisce eventualmente con le prestazioni e la stabilità.

Le zone MultiTarget sono generalmente considerate sicure, anche se in rare circostanze il comportamento delle porte target FC di fornitori diversi ha causato problemi. Ad esempio, evita di includere nella stessa zona le porte di destinazione di uno storage array NetApp e non NetApp. Inoltre, l'inserimento di un sistema di storage NetApp e di un dispositivo a nastro nella stessa zona è ancora più probabile che causino problemi.

Connessione di rete diretta

Gli amministratori dello storage a volte preferiscono semplificare le loro infrastrutture rimuovendo gli switch di rete dalla configurazione. Questo può essere supportato in alcuni scenari.

ISCSI e NVMe/TCP

Un host che utilizza iSCSI o NVMe/TCP può essere collegato direttamente a un sistema storage e funzionare normalmente. La ragione è la pedata. Le connessioni dirette a due storage controller differenti offrono due percorsi indipendenti per il flusso di dati. La perdita di percorso, porta o controller non impedisce l'utilizzo dell'altro percorso.

NFS

È possibile utilizzare lo storage NFS con connessione diretta, ma con una limitazione significativa: Il failover non funzionerà senza una significativa attività di scripting, che sarà responsabilità del cliente.

Il motivo per cui il failover senza interruzioni è complicato con lo storage NFS connesso direttamente è il routing che si verifica sul sistema operativo locale. Ad esempio, si supponga che un host abbia un indirizzo IP 192.168.1.1/24 e che sia collegato direttamente a un controller ONTAP con un indirizzo IP 192.168.1.50/24. Durante il failover, l'indirizzo 192.168.1.50 può eseguire il failover sull'altro controller e sarà disponibile per l'host, ma in che modo l'host rileva la sua presenza? L'indirizzo 192.168.1.1 originale esiste ancora sulla scheda di rete host che non si connette più a un sistema operativo. Il traffico destinato a 192.168.1.50 continuerebbe ad essere inviato a una porta di rete inutilizzabile.

La seconda scheda NIC del sistema operativo potrebbe essere configurata come 19 2.168.1.2 e sarebbe in grado di comunicare con l'indirizzo 192.168.1.50 non riuscito, ma le tabelle di routing locali avrebbero un valore predefinito di utilizzo di un solo indirizzo **e di un solo indirizzo** per comunicare con la subnet 192.168.1.0/24. Un amministratore di sistema potrebbe creare un framework di script che rilevi una connessione di rete non riuscita e alteri le tabelle di routing locali o che porti le interfacce verso l'alto e verso il basso. La procedura esatta dipende dal sistema operativo in uso.

In pratica, i clienti NetApp dispongono di NFS con connessione diretta, ma in genere solo per i workload in cui le pause io durante i failover sono accettabili. Quando si utilizzano i supporti rigidi, non devono verificarsi errori di i/o durante tali pause. L'io dovrebbe bloccarsi finché i servizi non vengono ripristinati, mediante failback o intervento manuale, per spostare gli indirizzi IP tra le schede NIC dell'host.

Connessione diretta FC

Non è possibile connettere direttamente un host a un sistema storage ONTAP utilizzando il protocollo FC. Il motivo è l'uso di NPIV. Il WWN che identifica una porta FC ONTAP per la rete FC utilizza un tipo di virtualizzazione chiamato NPIV. Qualsiasi dispositivo collegato a un sistema ONTAP deve essere in grado di riconoscere un WWN NPIV. Attualmente non vi sono fornitori di HBA che offrono un HBA che può essere installato in un host in grado di supportare un target NPIV.

Configurazione di rete sui sistemi ASA r2

Interfacce logiche

I database Oracle devono accedere allo storage. Le interfacce logiche (LIF) sono le tubazioni di rete che collegano una Storage Virtual Machine (SVM) alla rete e a loro volta

al database. La corretta progettazione della LIF è necessaria per garantire una larghezza di banda sufficiente per ogni carico di lavoro del database e il failover non comporta una perdita dei servizi storage.

Questa sezione fornisce una panoramica dei principi chiave di progettazione LIF per i sistemi ASA r2, ottimizzati per ambienti solo SAN. Per una documentazione più completa, vedere il "Documentazione di gestione della rete ONTAP". Come per altri aspetti dell'architettura del database, le migliori opzioni per la progettazione della macchina virtuale di archiviazione (SVM, nota come vserver nella CLI) e dell'interfaccia logica (LIF) dipendono in larga misura dai requisiti di scalabilità e dalle esigenze aziendali.

Durante la creazione di una strategia LIF, prendi in considerazione i seguenti argomenti principali:

- **Prestazione.** La larghezza di banda della rete è sufficiente per i carichi di lavoro Oracle?
- **Resilienza.** ci sono singoli punti di guasto nel progetto?
- **Gestibilità.** la rete può essere scalata senza interruzioni?

Gli argomenti trattati sono relativi alla soluzione end-to-end, dall'host fino agli switch fino al sistema storage.

Tipi di LIF

Esistono diversi tipi di LIF. "Documentazione ONTAP sui tipi di LIF" Fornisci informazioni più complete su questo argomento, ma da un punto di vista funzionale le LIF possono essere divise in gruppi:

- **LIF di gestione cluster e nodi.** LIF utilizzati per gestire il cluster storage.
- **LIF di gestione SVM.** interfacce che consentono l'accesso a una SVM tramite l'API REST o ONTAPI (nota anche come ZAPI) per funzioni come la creazione di snapshot o il ridimensionamento del volume. Prodotti come SnapManager for Oracle (SMO) devono avere accesso a una LIF di gestione SVM.
- **LIF dei dati.** Interfacce solo per protocolli SAN: FC, iSCSI, NVMe/FC, NVMe/TCP. I protocolli NAS (NFS, SMB/CIFS) non sono supportati sui sistemi ASA r2.

> Non è possibile configurare un'interfaccia sia per il traffico iSCSI (o NVMe/TCP) che per quello di gestione, nonostante entrambi utilizzino un protocollo IP. Negli ambienti iSCSI o NVMe/TCP è necessario un LIF di gestione separato. Per garantire resilienza e prestazioni, configurare più LIF di dati SAN per protocollo per nodo e distribuirli su diverse porte fisiche e fabric. A differenza dei sistemi AFF/ FAS , ASA r2 non consente il traffico NFS o SMB, quindi non è possibile riutilizzare un LIF dati NAS per la gestione.

Progettazione della SAN LIF

Il design di LIF in un ambiente SAN è relativamente semplice per un motivo: Il multipathing. Tutte le moderne implementazioni SAN consentono a un client di accedere ai dati su più percorsi di rete indipendenti e di selezionare i percorsi migliori per l'accesso. Di conseguenza, le performance rispetto alla progettazione LIF sono più semplici da gestire, perché i client SAN bilanciano automaticamente il carico dell'i/o nei migliori percorsi disponibili.

Se un percorso non è disponibile, il client seleziona automaticamente un percorso diverso. Grazie alla sua semplicità di progettazione, le LIF SAN sono generalmente più gestibili. Ciò non significa che un ambiente SAN sia sempre più facile da gestire, poiché vi sono molti altri aspetti dello storage SAN che sono molto più complicati di NFS. Significa semplicemente che la progettazione della SAN LIF è più semplice.

Performance

L'aspetto più importante da considerare per le prestazioni LIF in un ambiente SAN è la larghezza di banda. Ad esempio, un cluster ASA r2 a due nodi con due porte FC da 32 Gb per nodo consente fino a 64 Gb di larghezza di banda da/verso ciascun nodo. Allo stesso modo, per NVMe/TCP o iSCSI, assicurarsi di disporre di una connettività 25GbE o 100GbE sufficiente per i carichi di lavoro Oracle.

Resilienza

I sistemi SAN LIF non eseguono il failover nello stesso modo dei sistemi NAS LIF. I sistemi ASA r2 si basano sul multipathing host (MPIO/ALUA) per la resilienza. Se un SAN LIF diventa non disponibile a causa del failover del controller, il software multipathing del client rileva la perdita di un percorso e reindirizza l'I/O a un percorso alternativo. ASA r2 può eseguire la rilocazione LIF dopo un breve ritardo per ripristinare la piena disponibilità del percorso, ma ciò non interrompe l'I/O perché i percorsi attivi esistono già sul nodo partner. Il processo di failover si verifica per ripristinare l'accesso host su tutte le porte definite.

Gestibilità

Non è necessario migrare un LIF in un ambiente SAN quando i volumi vengono riposizionati all'interno della coppia HA. Questo perché, una volta completato lo spostamento del volume, ONTAP invia una notifica alla SAN in merito a una modifica nei percorsi e i client SAN eseguono automaticamente la riottimizzazione. La migrazione LIF con SAN è associata principalmente a importanti modifiche hardware fisiche. Ad esempio, se è necessario un aggiornamento non distruttivo dei controller, un SAN LIF viene migrato al nuovo hardware. Se una porta FC risulta difettosa, è possibile migrare un LIF su una porta non utilizzata.

Raccomandazioni di progettazione

NetApp fornisce le seguenti raccomandazioni per gli ambienti SAN ASA r2:

- Non creare più percorsi di quelli richiesti. Un numero eccessivo di percorsi complica la gestione complessiva e può causare problemi con il failover del percorso su alcuni host. Inoltre, alcuni host hanno limitazioni inattese del percorso per configurazioni come l'avvio SAN.
- Un numero molto ridotto di configurazioni deve richiedere più di quattro percorsi a un LUN. Il valore di avere più di due nodi che pubblicizzano i percorsi delle LUN è limitato perché l'aggregato che ospita un LUN è inaccessibile in caso di guasto del nodo proprietario del LUN e del partner ha. In una situazione del genere, la creazione di percorsi su nodi diversi dalla coppia ha primaria non è d'aiuto.
- Sebbene il numero di percorsi LUN visibili possa essere gestito selezionando le porte incluse nelle zone FC, in genere è più semplice includere tutti i potenziali punti di destinazione nella zona FC e controllare la visibilità delle LUN a livello ONTAP.
- Utilizzare la funzionalità di mappatura LUN selettiva (SLM), abilitata per impostazione predefinita. Con SLM, ogni nuova LUN viene automaticamente pubblicizzata dal nodo proprietario dell'aggregato sottostante e dal partner HA del nodo. Questa disposizione evita la necessità di creare set di porte o di configurare la suddivisione in zone per limitare l'accessibilità delle porte. Ogni LUN è disponibile sul numero minimo di nodi richiesti per garantire prestazioni e resilienza ottimali.
- Nel caso in cui una LUN debba essere migrata all'esterno dei due controller, i nodi aggiuntivi possono essere aggiunti con `lun mapping add-reporting-nodes` comando in modo che i LUN vengano pubblicizzati sui nuovi nodi. In questo modo si creano percorsi SAN aggiuntivi verso le LUN per la migrazione delle LUN. Tuttavia, l'host deve eseguire un'operazione di individuazione per utilizzare i nuovi percorsi.
- Non preoccupatevi eccessivamente del traffico indiretto. Si consiglia di evitare il traffico indiretto in un ambiente i/o-intensive per il quale è critico ogni microsecondo di latenza, ma l'effetto visibile delle performance è trascurabile per i workload tipici.

Configurazione TCP/IP ed ethernet

Molti clienti Oracle su ASA r2 ONTAP utilizzano Ethernet, il protocollo di rete di iSCSI e NVMe/TCP.

Impostazioni del sistema operativo host

La maggior parte della documentazione del fornitore di applicazioni include impostazioni TCP ed ethernet specifiche per garantire il funzionamento ottimale dell'applicazione. Queste stesse impostazioni sono in genere sufficienti per fornire anche prestazioni ottimali dello storage basato su IP.

Controllo di flusso Ethernet

Questa tecnologia consente a un client di richiedere che un mittente interrompa temporaneamente la trasmissione dei dati. Questa operazione viene solitamente eseguita perché il ricevitore non è in grado di elaborare i dati in ingresso abbastanza rapidamente. Una volta, la richiesta che un mittente cessi la trasmissione era meno disgregativa di avere pacchetti di scarto del destinatario perché i buffer erano pieni. Questo non è più il caso degli stack TCP utilizzati oggi nei sistemi operativi. Infatti, il controllo di flusso causa più problemi di quanti ne risolva.

Negli ultimi anni sono aumentati i problemi di prestazioni causati dal controllo di flusso Ethernet. Questo perché il controllo di flusso Ethernet opera al livello fisico. Se una configurazione di rete consente a qualsiasi sistema operativo host di inviare una richiesta di controllo di flusso Ethernet a un sistema di storage, il risultato è una pausa in i/o per tutti i client connessi. Poiché un numero crescente di client viene servito da un singolo storage controller, aumenta la probabilità che uno o più client inviino richieste di controllo di flusso. Il problema è stato riscontrato frequentemente presso le sedi dei clienti con un'ampia virtualizzazione del sistema operativo.

Una scheda NIC su un sistema NetApp non dovrebbe ricevere richieste di controllo di flusso. Il metodo utilizzato per ottenere questo risultato varia in base al produttore dello switch di rete. Nella maggior parte dei casi, il controllo di flusso su uno switch Ethernet può essere impostato su `receive desired` oppure `receive on`, il che significa che una richiesta di controllo di flusso non viene inoltrata al controller di memorizzazione. In altri casi, la connessione di rete sul controller di storage potrebbe non consentire la disattivazione del controllo di flusso. In questi casi, i client devono essere configurati in modo da non inviare mai richieste di controllo di flusso, modificando la configurazione NIC sul server host stesso o le porte switch a cui è connesso il server host.

Per i sistemi ASA r2, che sono solo SAN, le considerazioni sul controllo del flusso Ethernet si applicano principalmente al traffico iSCSI e NVMe/TCP.

* NetApp consiglia* di assicurarsi che i controller di storage NetApp ASA r2 non ricevano pacchetti di controllo del flusso Ethernet. In genere, questa operazione può essere eseguita impostando le porte dello switch a cui è collegato il controller, ma alcuni hardware dello switch presentano delle limitazioni che potrebbero richiedere modifiche lato client.

Dimensioni MTU

È stato dimostrato che l'utilizzo dei frame jumbo offre un certo miglioramento delle performance nelle reti 1Gb, riducendo l'overhead della CPU e della rete, ma i benefici non sono solitamente significativi.

NetApp consiglia l'implementazione di frame jumbo quando possibile, sia per ottenere potenziali vantaggi in termini di prestazioni sia per rendere la soluzione a prova di futuro.

Per i sistemi ASA r2, che sono solo SAN, i frame jumbo si applicano solo ai protocolli SAN basati su Ethernet (iSCSI e NVMe/TCP).

L'utilizzo di frame jumbo in una rete 10Gb è quasi obbligatorio. Questo perché la maggior parte delle implementazioni 10Gb raggiungono un limite di pacchetti al secondo senza frame jumbo prima che raggiungano il contrassegno 10Gb. L'utilizzo di frame jumbo migliora l'efficienza dell'elaborazione TCP/IP, poiché consente a sistema operativo, server, schede di rete e sistema di storage di elaborare un numero inferiore di pacchetti, anche se di dimensioni maggiori. Il miglioramento delle prestazioni varia da scheda di rete a scheda di rete, ma è significativo.

Per le implementazioni jumbo-frame, esiste la convinzione comune, ma non corretta, che tutti i dispositivi connessi debbano supportare frame jumbo e che le dimensioni MTU debbano corrispondere end-to-end Al contrario, i due endpoint di rete negoziano la dimensione del frame più elevata reciprocamente accettabile quando si stabilisce una connessione. In un ambiente tipico, uno switch di rete è impostato su una dimensione MTU di 9216, il controller NetApp è impostato su 9000 e i client sono impostati su una combinazione di 9000 e 1514. I client in grado di supportare un valore MTU di 9000 possono utilizzare frame jumbo, mentre i client in grado di supportare solo 1514 possono negoziare un valore inferiore.

I problemi con questa disposizione sono rari in un ambiente completamente commutato. Tuttavia, in un ambiente con routing occorre assicurarsi che nessun router intermedio sia costretto a frammentare frame jumbo.

- NetApp consiglia* di configurare quanto segue per gli ambienti SAN ASA r2:
- I jumbo frame sono desiderabili ma non obbligatori con 1 GbE.
- Per ottenere le massime prestazioni con 10 GbE sono necessari i frame jumbo e velocità superiori per il traffico iSCSI e NVMe/TCP.

Parametri TCP

Tre impostazioni spesso non sono configurate correttamente: Timestamp TCP, riconoscimento selettivo (SACK) e ridimensionamento finestra TCP. Molti documenti obsoleti su Internet consigliano di disabilitare uno o più di questi parametri per migliorare le prestazioni. Molti anni fa, questa raccomandazione ha avuto un certo merito quando le capacità della CPU erano molto inferiori e, quando possibile, vi era un vantaggio nel ridurre il sovraccarico sull'elaborazione TCP.

Tuttavia, con i sistemi operativi moderni, la disattivazione di una qualsiasi di queste funzioni TCP in genere non comporta alcun vantaggio rilevabile e, allo stesso tempo, può danneggiare le prestazioni. In ambienti di rete virtualizzati, i danni alle prestazioni sono particolarmente probabili, poiché queste funzioni sono necessarie per gestire in modo efficiente la perdita di pacchetti e le modifiche della qualità della rete.

NetApp consiglia di abilitare timestamp TCP, SACCO e ridimensionamento finestra TCP sull'host, e tutti e tre questi parametri dovrebbero essere attivi per impostazione predefinita in qualsiasi sistema operativo corrente.

Configurazione FC SAN

La configurazione di FC SAN per database Oracle su sistemi ASA r2 consiste principalmente nel seguire le best practice SAN standard.

ASA r2 è ottimizzato per carichi di lavoro esclusivamente SAN, pertanto i principi rimangono gli stessi di AFF/FAS, con particolare attenzione a prestazioni, resilienza e semplicità. Ciò include misure di pianificazione tipiche, come la garanzia che esista una larghezza di banda sufficiente sulla SAN tra l'host e il sistema di

archiviazione, la verifica che tutti i percorsi SAN esistano tra tutti i dispositivi richiesti, l'utilizzo delle impostazioni della porta FC richieste dal fornitore dello switch FC, l'evitamento di contese ISL e l'utilizzo di un monitoraggio adeguato della struttura SAN.

Suddivisione in zone

Una zona FC non deve mai contenere più di un iniziatore. Una tale disposizione potrebbe sembrare funzionare inizialmente, ma la diafonia tra gli iniziatori interferisce eventualmente con le prestazioni e la stabilità.

Le zone MultiTarget sono generalmente considerate sicure, anche se in rare circostanze il comportamento delle porte target FC di fornitori diversi ha causato problemi. Ad esempio, evita di includere nella stessa zona le porte di destinazione di uno storage array NetApp e non NetApp. Inoltre, l'inserimento di un sistema di storage NetApp e di un dispositivo a nastro nella stessa zona è ancora più probabile che causino problemi.

- ASA r2 utilizza le zone di disponibilità dello storage anziché gli aggregati, ma ciò non modifica i principi di zonizzazione FC.
- Il multipathing (MPIO) rimane il meccanismo di resilienza principale; tuttavia, per i sistemi ASA r2 che supportano il multipathing attivo-attivo simmetrico, tutti i percorsi verso una LUN sono attivi e utilizzati simultaneamente per l'I/O.

Connessione di rete diretta

Gli amministratori dello storage a volte preferiscono semplificare le loro infrastrutture rimuovendo gli switch di rete dalla configurazione. Questo può essere supportato in alcuni scenari.

ISCSI e NVMe/TCP

Un host che utilizza iSCSI o NVMe/TCP può essere collegato direttamente a un sistema di archiviazione ASA r2 e funzionare normalmente. Il motivo è il pathing. Le connessioni dirette a due diversi controller di archiviazione danno luogo a due percorsi indipendenti per il flusso di dati. La perdita di un percorso, di una porta o di un controller non impedisce l'utilizzo dell'altro percorso, a condizione che il multipathing sia configurato correttamente.

Connessione diretta FC

Non è possibile connettere direttamente un host a un sistema di archiviazione ASA r2 utilizzando il protocollo FC. Il motivo è lo stesso dei sistemi AFF/ FAS : l'uso dell'NPIV. Il WWN che identifica una porta ONTAP FC nella rete FC utilizza un tipo di virtualizzazione denominato NPIV. Qualsiasi dispositivo connesso a un sistema ONTAP deve essere in grado di riconoscere un NPIV WWN. Attualmente non ci sono fornitori di HBA che offrano un HBA installabile in un host in grado di supportare un target NPIV.

Configurazione dello storage su sistemi AFF/FAS

SAN FC

Allineamento delle LUN

L'allineamento delle LUN si riferisce all'ottimizzazione dell'i/o in relazione al layout del file system sottostante.

Su un sistema ONTAP, lo storage è organizzato in 4KB unità. Un blocco 8KB di un database o di un file system deve corrispondere esattamente a due blocchi 4KB. Se un errore nella configurazione LUN sposta l'allineamento di 1KB:1 in entrambe le direzioni, ogni blocco 8KB esisterebbe su tre blocchi di storage 4KB diversi invece che due. Questa disposizione causerebbe un aumento della latenza e causerebbe l'esecuzione di ulteriori i/o all'interno del sistema di storage.

L'allineamento influisce anche sulle architetture LVM. Se un volume fisico all'interno di un gruppo di volumi logici viene definito sull'intero dispositivo del disco (non vengono create partizioni), il primo blocco 4KB sul LUN si allinea con il primo blocco 4KB sul sistema di storage. Questo è un allineamento corretto. I problemi si verificano con le partizioni perché spostano la posizione iniziale in cui il sistema operativo utilizza il LUN. Finché l'offset viene spostato in intere unità di 4KB, il LUN viene allineato.

Negli ambienti Linux, creare gruppi di volumi logici sull'intero dispositivo di unità. Quando è necessaria una partizione, controllare l'allineamento eseguendo `fdisk -u` e verificando che l'inizio di ogni partizione sia un multiplo di otto. Ciò significa che la partizione inizia da un multiplo di otto settori a 512 byte, ovvero 4KB.

Vedere anche la discussione sull'allineamento dei blocchi di compressione nella sezione "Efficienza". Qualsiasi layout allineato ai limiti del blocco di compressione 8KB è allineato ai limiti 4KB.

Avvertenze di disallineamento

La registrazione di ripristino del database/transazioni genera di solito un i/o non allineato che può causare avvisi fuorvianti riguardo ai LUN disallineati su ONTAP.

La registrazione esegue una scrittura sequenziale del file di registro con scritture di dimensioni variabili. Un'operazione di scrittura del registro che non si allinea ai limiti 4KB non causa normalmente problemi di prestazioni poiché l'operazione di scrittura del registro successiva completa il blocco. Il risultato è che ONTAP è in grado di elaborare quasi tutte le scritture come blocchi da 4KB KB completi, anche se i dati in alcuni blocchi da 4KB KB sono stati scritti in due operazioni separate.

Verificare l'allineamento utilizzando utilità come `sio` oppure `dd` Che possono generare i/o a dimensioni dei blocchi definite. È possibile visualizzare le statistiche di allineamento di i/o del sistema di storage con `stats` comando. Vedere "Verifica dell'allineamento di WAFL" per ulteriori informazioni.

L'allineamento negli ambienti Solaris è più complicato. Fare riferimento a. "Configurazione host SAN ONTAP" per ulteriori informazioni.

Attenzione
Negli ambienti Solaris x86, prestare ulteriore attenzione al corretto allineamento poiché la maggior parte delle configurazioni prevede diversi livelli di partizioni. Le sezioni di partizione di Solaris x86 si trovano solitamente in cima a una tabella di partizioni del record di avvio master standard.

Dimensionamento e numero di LUN

La scelta delle dimensioni ottimali e del numero di LUN da utilizzare è un elemento critico per ottenere performance e gestibilità ottimali dei database Oracle.

Un LUN è un oggetto virtualizzato in ONTAP presente in tutti i dischi dell'aggregato di hosting. Di conseguenza, le performance della LUN non sono influenzate dalle sue dimensioni, perché la LUN sfrutta al massimo il potenziale in termini di performance dell'aggregato, indipendentemente dalle dimensioni scelte.

Per comodità, i clienti potrebbero desiderare di utilizzare un LUN di particolari dimensioni. Ad esempio, se un database è costruito su un gruppo di dischi LVM o Oracle ASM composto da due LUN da 1TB GB ciascuno,

tale gruppo di dischi deve essere aumentato in incrementi di 1TB TB. Potrebbe essere preferibile costruire il gruppo di dischi da otto LUN da 500GB ciascuno in modo che il gruppo di dischi possa essere aumentato con incrementi più piccoli.

La pratica di stabilire una dimensione LUN standard universale è scoraggiata perché ciò può complicare la gestibilità. Ad esempio, è possibile che una dimensione LUN standard di 100GB TB sia ottimale quando un database o un datastore è compreso nell'intervallo da 1TB a 2TB TB, ma un database o un datastore di 20TB GB richiederebbe 200 LUN. Ciò significa che i tempi di riavvio del server sono più lunghi, che vi sono più oggetti da gestire nelle varie interfacce utente e che prodotti come SnapCenter devono eseguire la ricerca su molti oggetti. Utilizzando un numero inferiore di LUN di dimensioni maggiori è possibile evitare questi problemi.

- Il numero di LUN è più importante delle dimensioni delle LUN.
- Le dimensioni dei LUN sono principalmente controllate dai requisiti di numero di LUN.
- Evitare di creare più LUN del necessario.

Numero di LUN

A differenza delle dimensioni delle LUN, il numero di LUN influisce sulle performance. Spesso le prestazioni delle applicazioni dipendono dalla capacità di eseguire i/o paralleli attraverso il livello SCSI. Di conseguenza, due LUN offrono performance migliori rispetto a una singola LUN. Utilizzare un LVM come Veritas VxVM, Linux LVM2 o Oracle ASM è il metodo più semplice per aumentare il parallelismo.

I clienti di NetApp hanno in genere ottenuto il minimo beneficio dall'aumento del numero di LUN oltre i sedici, sebbene i test degli ambienti con dischi a stato solido al 100% con i/o casuali molto intensi abbiano dimostrato un ulteriore miglioramento fino a 64 LUN.

NetApp consiglia quanto segue:

In generale, da quattro a sedici LUN sono sufficienti per supportare le esigenze di i/o di qualsiasi carico di lavoro del database. Meno di quattro LUN potrebbero creare limiti di performance a causa delle limitazioni nelle implementazioni SCSI host.

Posizionamento delle LUN

Il posizionamento ottimale delle LUN del database all'interno dei volumi ONTAP dipende principalmente dalle diverse funzionalità di ONTAP.

Volumi

Un punto comune di confusione tra i clienti che non conoscono ONTAP è l'utilizzo di FlexVol, comunemente denominati semplicemente "volumi".

Un volume non è un LUN. Questi termini vengono utilizzati in maniera anonima con molti prodotti di altri vendor, inclusi i cloud provider. ONTAP Volumes sono semplicemente container di gestione. Non forniscono dati da soli, né occupano spazio. Sono container per file o LUN e esistono per migliorare e semplificare la gestibilità, in particolare su larga scala.

Volumi e LUN

I LUN correlati sono normalmente collocati in una stessa posizione in un singolo volume. Ad esempio, un database che richiede 10 LUN solitamente conterrà tutte le 10 LUN dello stesso volume.

- L'utilizzo di un rapporto di 1:1:1 tra LUN e volumi, vale a dire un LUN per volume, non è * una Best practice formale.
- I volumi dovrebbero invece essere visti come container per i carichi di lavoro o i set di dati. È possibile che sia presente un singolo LUN per volume o che ve ne siano molti. La risposta giusta dipende dai requisiti di gestibilità.
- La dispersione dei LUN in un numero non necessario di volumi può causare overhead e problemi di pianificazione aggiuntivi per operazioni quali operazioni di snapshot, un numero eccessivo di oggetti visualizzati nell'interfaccia utente e il raggiungimento dei limiti di volume della piattaforma prima del raggiungimento del limite LUN.

Volumi, LUN e snapshot

I criteri e le pianificazioni degli Snapshot vengono posizionati sul volume, non sul LUN. Un set di dati composto da 10 LUN richiederebbe una singola policy di snapshot quando le LUN sono collocate contemporaneamente nello stesso volume.

Inoltre, la co-localizzazione di tutti i LUN correlati per un dato dataset in un singolo volume consente di eseguire operazioni di snapshot atomiche. Ad esempio, un database di 10 LUN o un ambiente applicativo basato su VMware composto da 10 diversi sistemi operativi possono essere protetti come un singolo oggetto coerente se le LUN sottostanti vengono tutte collocate in un singolo volume. Se vengono posizionati su volumi diversi, gli snapshot possono essere o meno sincronizzati al 100%, anche se pianificati allo stesso tempo.

In alcuni casi, potrebbe essere necessario suddividere una serie di LUN correlata in due volumi diversi a causa dei requisiti di recovery. Ad esempio, un database potrebbe avere quattro LUN per i file di dati e due LUN per i log. In questo caso, un volume di file dati con 4 LUN e un volume di registro con 2 LUN potrebbe essere l'opzione migliore. Il motivo è la possibilità di recupero indipendente. Ad esempio, è possibile ripristinare in maniera selettiva il volume di file dati a uno stato precedente, vale a dire che le quattro LUN vengono riportate allo stato della snapshot, senza influire sul volume di log con i dati critici.

Volumi, LUN e SnapMirror

Le policy e le operazioni di SnapMirror, come le operazioni di Snapshot, vengono eseguite sul volume, non sul LUN.

La co-localizzazione dei LUN correlati in un singolo volume consente di creare una singola relazione di SnapMirror e di aggiornare tutti i dati contenuti con un singolo update. Come per gli snapshot, l'aggiornamento sarà anche un'operazione atomica. La destinazione SnapMirror avrà una replica point-in-time singola delle LUN di origine. Se le LUN sono state distribuite su più volumi, le repliche possono essere o meno coerenti l'una con l'altra.

Volumi, LUN e QoS

Mentre la qualità del servizio può essere applicata in modo selettivo alle singole LUN, in genere è più semplice impostarla a livello di volume. Ad esempio, tutte le LUN utilizzate dai guest di un determinato server ESX possono essere collocate su un singolo volume e successivamente può essere applicata una policy di QoS adattiva di ONTAP. In questo modo si ottiene un limite di IOPS per TB autoscalabile valido per tutte le LUN.

Analogamente, se un database richiedeva 100K IOPS e occupava 10 LUN, sarebbe più semplice impostare un singolo limite di 100K IOPS su un singolo volume piuttosto che impostare 10 limiti individuali di 10K IOPS, uno per ogni LUN.

Layout a più volumi

Vi sono alcuni casi in cui la distribuzione delle LUN su più volumi può essere vantaggiosa. Il motivo principale è lo striping dei controller. Ad esempio, un sistema storage ha potrebbe ospitare un singolo database in cui è richiesto il potenziale completo di elaborazione e caching di ogni controller. In questo caso, una progettazione tipica sarebbe quella di collocare metà dei LUN in un singolo volume sul controller 1, e l'altra metà dei LUN in un singolo volume sul controller 2.

Analogamente, lo striping dei controller potrebbe essere utilizzato per il bilanciamento del carico. Un sistema ha che ospitava 100 database da 10 LUN ciascuno potrebbe essere progettato dove ogni database riceve un volume da 5 LUN su ciascuno dei due controller. Il risultato è garantito il caricamento simmetrico di ogni controller quando vengono forniti database aggiuntivi.

Tuttavia, nessuno di questi esempi riguarda un rapporto volume-LUN di 1:1:1. L'obiettivo resta quello di ottimizzare la gestibilità mediante la co-localizzazione dei LUN correlati in volumi.

Un esempio se un rapporto da 1:1 LUN a volume è sensato è la containerizzazione, laddove ogni LUN potrebbe rappresentare davvero un singolo carico di lavoro e deve essere gestita singolarmente. In questi casi, un rapporto 1:1:1 può essere ottimale.

Ridimensionamento LUN e ridimensionamento LVM

Quando un file system basato su SAN ha raggiunto il limite di capacità, sono disponibili due opzioni per aumentare lo spazio disponibile:

- Aumentare la dimensione dei LUN
- Aggiungere un LUN a un gruppo di volumi esistente e aumentare il volume logico contenuto

Sebbene il ridimensionamento delle LUN sia un'opzione per aumentare la capacità, in genere è preferibile utilizzare un LVM, incluso Oracle ASM. Uno dei motivi principali per cui esistono le LVM è evitare la necessità di ridimensionare le LUN. Con un LVM, più LUN sono unite in un pool virtuale di storage. I volumi logici scavati da questo pool sono gestiti da LVM e possono essere facilmente ridimensionati. Un ulteriore vantaggio è l'eliminazione degli hotspot su una determinata unità distribuendo un determinato volume logico su tutte le LUN disponibili. Di solito, la migrazione trasparente può essere eseguita utilizzando il volume manager per spostare le estensioni sottostanti di un volume logico su nuovi LUN.

Striping LVM

Lo striping LVM si riferisce alla distribuzione dei dati su più LUN. Il risultato è un significativo miglioramento delle performance per molti database.

Prima dell'era dei dischi flash, era stato utilizzato lo striping per superare i limiti di performance dei dischi rotanti. Ad esempio, se un sistema operativo deve eseguire un'operazione di lettura a 1MB bit, la lettura di 1MB GB di dati da un'unica unità richiederebbe un'ampia ricerca e lettura della testina dell'unità poiché il sistema 1MB viene trasferito lentamente. Se quei 1MB TB di dati sono stati suddivisi in 8 LUN, il sistema operativo potrebbe emettere otto operazioni di lettura 128K in parallelo, riducendo il tempo necessario per completare il trasferimento da 1MB GB.

Lo striping con dischi rotanti era più difficile perché lo schema di i/o doveva essere noto in anticipo. Se lo striping non è stato regolato correttamente per i modelli i/o reali, le configurazioni con striping potrebbero danneggiare le prestazioni. Con i database Oracle, e in particolare con le configurazioni all-flash, lo striping è molto più semplice da configurare ed è stato dimostrato che le performance risultano notevolmente migliorate.

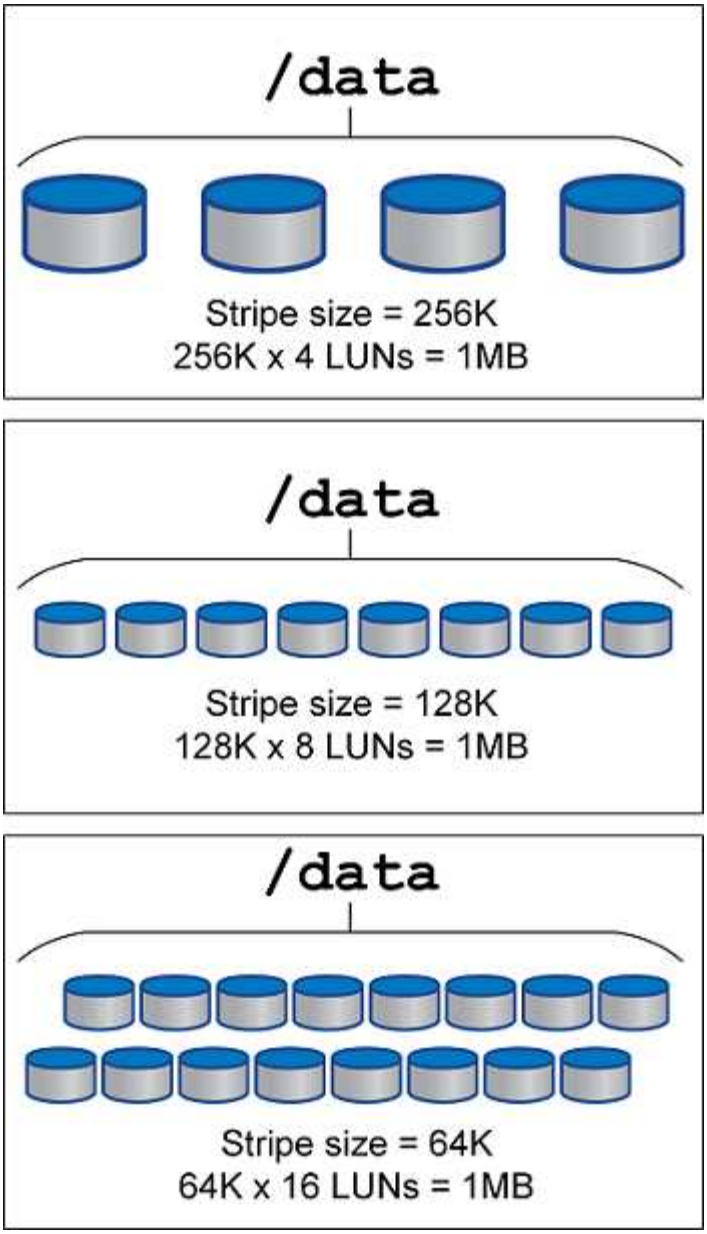
Per impostazione predefinita, i gestori di volume logici, come lo stripe di Oracle ASM, ma il sistema operativo

LVM nativo non lo fanno. Alcune di esse collegano più LUN insieme come un dispositivo concatenato, il che comporta file di dati che esistono su un solo dispositivo LUN. Ciò causa punti caldi. Le altre implementazioni LVM sono impostate per impostazione predefinita su estensioni distribuite. Questo è simile allo striping, ma è più grossolano. I LUN nel gruppo di volumi vengono suddivisi in porzioni di grandi dimensioni, chiamate estensioni e generalmente misurati in molti megabyte, e i volumi logici vengono quindi distribuiti tra tali estensioni. Il risultato è un i/o casuale per un file dovrebbe essere ben distribuito tra i LUN, ma le operazioni i/o sequenziali non sono così efficienti come potrebbero essere.

L'i/o delle applicazioni che richiedono elevate performance è quasi sempre (a) in unità delle dimensioni dei blocchi di base o (b) un megabyte.

L'obiettivo principale di una configurazione con striping è quello di garantire che l'i/o a file singolo possa essere eseguito come una singola unità, mentre l'i/o a blocchi multipli, di dimensioni pari a 1MB GB, può essere parallelizzato in modo uniforme tra tutti i LUN del volume con striping. Ciò significa che la dimensione dello stripe non deve essere inferiore alla dimensione del blocco del database e che la dimensione dello stripe moltiplicata per il numero di LUN deve essere 1MB.

La figura seguente mostra tre possibili opzioni per la regolazione delle dimensioni dello stripe e della larghezza. Il numero di LUN viene selezionato per soddisfare i requisiti di prestazioni come descritto sopra, ma in tutti i casi i dati totali all'interno di uno stripe singolo sono 1MB.



NFS

Panoramica

NetApp offre storage NFS Enterprise da oltre 30 anni e il suo utilizzo cresce insieme alla spinta verso infrastrutture basate sul cloud grazie alla sua semplicità.

Il protocollo NFS include diverse versioni con diversi requisiti. Per una descrizione completa della configurazione NFS con ONTAP, vedere "Best practice NFS su ONTAP TR-4067". Le sezioni seguenti descrivono alcuni dei requisiti più critici e gli errori comuni degli utenti.

Versioni di NFS

Il client NFS del sistema operativo deve essere supportato da NetApp.

- NFSv3 è supportato con sistemi operativi che seguono lo standard NFSv3.

- NFSv3 è supportato con il client Oracle DNFS.
- NFSv4 è supportato con tutti i sistemi operativi che seguono lo standard NFSv4.
- I sistemi NFSv4,1 e NFSv4,2 richiedono supporto specifico per il sistema operativo. Consultare "NetApp IMT" Per i sistemi operativi supportati.
- Il supporto di Oracle DNFS per NFSv4,1 richiede Oracle 12.2.0.2 o versione successiva.

> Il "Matrice di supporto di NetApp" Per NFSv3 e NFSv4 non sono inclusi sistemi operativi specifici. Tutti i sistemi operativi che rispettano la RFC sono generalmente supportati. Quando si cerca il supporto NFSv3 o NFSv4 nel IMT online, non selezionare un sistema operativo specifico perché non verranno visualizzate corrispondenze. Tutti i sistemi operativi sono implicitamente supportati dalla policy generale.

Tabelle degli slot TCP per Linux NFSv3

Le tabelle degli slot TCP sono l'equivalente di NFSv3 della profondità della coda degli HBA (host Bus Adapter). Queste tabelle controllano il numero di operazioni NFS che possono essere in sospeso in qualsiasi momento. Il valore predefinito è di solito 16, che è troppo basso per ottenere prestazioni ottimali. Il problema opposto si verifica sui kernel Linux più recenti, che possono aumentare automaticamente il limite della tabella degli slot TCP a un livello che satura il server NFS con le richieste.

Per prestazioni ottimali e per evitare problemi di prestazioni, regolare i parametri del kernel che controllano le tabelle degli slot TCP.

Eseguire `sysctl -a | grep tcp.*.slot_table` e osservare i seguenti parametri:

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

Tutti i sistemi Linux dovrebbero includere `sunrpc.tcp_slot_table_entries`, ma solo alcuni includono `sunrpc.tcp_max_slot_table_entries`. Entrambi devono essere impostati su 128.

> La mancata impostazione di questi parametri può avere effetti significativi sulle prestazioni. In alcuni casi, le prestazioni sono limitate poiché il sistema operativo linux non fornisce i/o sufficienti In altri casi, le latenze i/o aumentano quando il sistema operativo linux tenta di emettere più i/o di quanto possa essere gestito.

ADR e NFS

Alcuni clienti hanno segnalato problemi di prestazioni derivanti da una quantità eccessiva di i/o sui dati in `ADR` posizione. Il problema generalmente non si verifica finché non si sono accumulati molti dati sulle prestazioni. Il motivo dell'eccessivo i/o è sconosciuto, ma questo problema sembra essere dovuto ai processi Oracle che eseguono ripetutamente la scansione della directory di destinazione per rilevare eventuali modifiche.

Smontaggio del `noac` e/o. `actimeo=0` Le opzioni di montaggio consentono il caching del sistema operativo host e riducono i livelli di i/o dello storage.

NetApp consiglia di non piazzare `ADR` dati su un file system con `noac` oppure `actimeo=0` perché è probabile che si verifichino problemi di prestazioni. Separare `ADR` dati in un punto di montaggio diverso, se necessario.

nfs-rootonly e mount-rootonly

ONTAP include un'opzione NFS denominata `nfs-rootonly` Che controlla se il server accetta connessioni di traffico NFS da porte elevate. Come misura di sicurezza, solo l'utente root è autorizzato ad aprire connessioni TCP/IP utilizzando una porta di origine inferiore a 1024, poiché tali porte sono normalmente riservate all'uso del sistema operativo, non ai processi utente. Questa restrizione aiuta a garantire che il traffico NFS provenga da un client NFS del sistema operativo effettivo e non da un processo dannoso che emula un client NFS. Il client Oracle DNFS è un driver userspace, ma il processo viene eseguito come root, quindi in genere non è necessario modificare il valore di `nfs-rootonly`. I collegamenti sono costituiti da porte basse.

Il `mount-rootonly` L'opzione è valida solo per NFSv3. Controlla se la chiamata di MONTAGGIO RPC può essere accettata dalle porte superiori a 1024. Quando si utilizza DNFS, il client viene nuovamente eseguito come root, in modo da poter aprire le porte al di sotto di 1024. Questo parametro non ha alcun effetto.

I processi che aprono connessioni con DNFS su NFS versione 4,0 e successive non vengono eseguiti come root e quindi richiedono porte su 1024. Il `nfs-rootonly` Il parametro deve essere impostato su disabilitato affinché DNFS completi la connessione.

Se `nfs-rootonly` È attivato, il risultato è un blocco durante la fase di mount che apre le connessioni DNFS. L'output di sqlplus è simile a questo:

```
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
```

Il parametro può essere modificato come segue:

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

In situazioni rare, potrebbe essere necessario modificare sia nfs-rootonly che mount-rootonly in disabled. Se un server gestisce un numero estremamente elevato di connessioni TCP, è possibile che non siano disponibili porte al di sotto di 1024 e che il sistema operativo sia costretto a utilizzare porte più elevate. Questi due parametri ONTAP devono essere modificati per consentire il completamento della connessione.

Policy di esportazione NFS: Superuser e setuid

Se i file binari Oracle si trovano in una condivisione NFS, la policy di esportazione deve includere autorizzazioni superser e setuid.

Le esportazioni NFS condivise utilizzate per servizi file generici come le home directory dell'utente spesso

fanno uso dell'utente root. Ciò significa che una richiesta da parte dell'utente root su un host che ha montato un filesystem viene rimappata come un altro utente con privilegi inferiori. In questo modo è possibile proteggere i dati impedendo a un utente root di un determinato server di accedere ai dati del server condiviso. Il bit setuid può anche essere un rischio per la protezione in un ambiente condiviso. Il bit setuid consente di eseguire un processo come un utente diverso da quello che richiama il comando. Ad esempio, uno script della shell di proprietà di root con il bit setuid viene eseguito come root. Se lo script della shell potrebbe essere modificato da altri utenti, qualsiasi utente non root potrebbe eseguire un comando come root aggiornando lo script.

I file binari di Oracle includono file di proprietà di root e utilizzano il bit setuid. Se i file binari Oracle sono installati su una condivisione NFS, la policy di esportazione deve includere le autorizzazioni appropriate per superutente e setuid. Nell'esempio seguente, la regola include entrambi `allow-suid` e permessi `superuser` Accesso root per client NFS utilizzando l'autenticazione di sistema.

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

Configurazione NFSv4/4,1

Per la maggior parte delle applicazioni, la differenza tra NFSv3 e NFSv4 è minima. L'i/o delle applicazioni è di solito un i/o molto semplice e non trae alcun vantaggio significativo da alcune delle funzionalità avanzate disponibili in NFSv4. Le versioni più elevate di NFS non devono essere considerate come un "aggiornamento" dal punto di vista dello storage dei database, ma come versioni di NFS che includono funzionalità aggiuntive. Ad esempio, se è richiesta la protezione end-to-end della modalità di privacy Kerberos (krb5p), è necessario NFSv4.

NetApp consiglia di utilizzare NFSv4,1 se sono necessarie funzionalità NFSv4. Sono stati apportati alcuni miglioramenti funzionali al protocollo NFSv4 di NFSv4,1 che migliorano la resilienza in alcuni casi edge.

Il passaggio a NFSv4 è più complicato che cambiare semplicemente le opzioni di montaggio da vers=3 a vers=4,1. Una spiegazione più completa della configurazione NFSv4 con ONTAP, incluse le istruzioni sulla configurazione del sistema operativo, vedere "Best practice TR-4067 NFS su ONTAP". Le seguenti sezioni di questo TR spiegano alcuni dei requisiti di base per l'utilizzo di NFSv4.

Dominio NFSv4

Una spiegazione completa della configurazione NFSv4/4,1 esula dall'ambito di questo documento, ma un problema comunemente riscontrato è una mancata corrispondenza nella mappatura del dominio. Dal punto di vista di sysadmin, i file system NFS sembrano comportarsi normalmente, ma le applicazioni segnalano errori relativi ai permessi e/o setuid su determinati file. In alcuni casi, gli amministratori hanno concluso erroneamente che le autorizzazioni dei binari dell'applicazione sono state danneggiate e hanno eseguito comandi chown o chmod quando il problema effettivo era il nome di dominio.

Il nome di dominio NFSv4 viene impostato sulla SVM ONTAP:

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

Il nome di dominio NFSv4 sull'host è impostato in `/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
```

I nomi di dominio devono corrispondere. In caso contrario, vengono visualizzati errori di mappatura simili a quelli riportati di seguito nella `/var/log/messages`:

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

I file binari delle applicazioni, come i file binari dei database Oracle, includono i file di proprietà di root con il bit setuid, il che significa che una mancata corrispondenza nei nomi di dominio NFSv4 causa errori nell'avvio di Oracle e un avviso sulla proprietà o sulle autorizzazioni di un file chiamato `oradism`, che si trova nella `$ORACLE_HOME/bin` directory. Dovrebbe comparire come segue:

```
[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
```

Se questo file viene visualizzato con proprietà di nessuno, potrebbe esserci un problema di mappatura del dominio NFSv4.

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

Per risolvere questo problema, controllare `/etc/idmap.cfg` Eseguire il file in base all'impostazione del dominio id v4 in ONTAP e assicurarsi che siano coerenti. In caso contrario, apportare le modifiche necessarie, eseguire `nfsidmap -c`, e attendere un momento per la propagazione delle modifiche. La proprietà del file dovrebbe quindi essere riconosciuta correttamente come root. Se un utente aveva tentato di eseguire `chown root` Su questo file prima che la configurazione dei domini NFS sia stata corretta, potrebbe essere necessario eseguire `chown root` di nuovo.

Oracle Direct NFS (DNFS)

I database Oracle possono utilizzare NFS in due modi.

In primo luogo, può usare un filesystem montato usando il client NFS nativo che fa parte del sistema operativo. Questo è talvolta chiamato kernel NFS, o kNFS. Il filesystem NFS è montato e usato dal database Oracle esattamente come qualsiasi altra applicazione userebbe un filesystem NFS.

Il secondo metodo è Oracle Direct NFS (DNFS). Si tratta di un'implementazione dello standard NFS nel software di database Oracle. Senza modificare le modalità di configurazione o gestione dei database Oracle da parte del DBA. Purché le impostazioni del sistema storage siano corrette, l'utilizzo del DNFS deve essere trasparente per il team DBA e gli utenti finali.

Un database con la funzione DNFS attivata ha ancora i consueti filesystem NFS montati. Una volta aperto il database, il database Oracle apre una serie di sessioni TCP/IP ed esegue direttamente le operazioni NFS.

NFS diretto

Il valore principale di Oracle Direct NFS è quello di ignorare il client NFS host ed eseguire operazioni di file NFS direttamente su un server NFS. Per abilitarla è sufficiente modificare la libreria Oracle Disk Manager (ODM). Le istruzioni per questo processo sono fornite nella documentazione di Oracle.

L'utilizzo di DNFS porta a un significativo miglioramento delle performance di i/o e riduce il carico sull'host e sul sistema storage poiché l'i/o viene eseguito nel modo più efficiente possibile.

Inoltre, Oracle DNFS include un'opzione **opzionale** per il multipathing e la fault tolerance dell'interfaccia di rete. Ad esempio, è possibile associare due interfacce 10Gb in modo da ottenere una larghezza di banda di 20Gb Gbps. Un errore di un'interfaccia provoca il tentativo di i/o sull'altra interfaccia. Il funzionamento complessivo è molto simile al multipathing FC. Il multipathing era comune anni fa quando ethernet a 1Gb GB rappresentava lo standard più comune. Una NIC 10Gb è sufficiente per la maggior parte dei carichi di lavoro Oracle, ma se ne richiede di più, è possibile collegare 10Gb NIC.

Quando si utilizza DNFS, è fondamentale che tutte le patch descritte in Oracle Doc 1495104,1 siano installate. Se non è possibile installare una patch, è necessario valutare l'ambiente per assicurarsi che i bug descritti in quel documento non causino problemi. In alcuni casi, l'impossibilità di installare le patch necessarie impedisce l'utilizzo di DNFS.

Non utilizzare DNFS con alcun tipo di risoluzione dei nomi round-robin, compresi DNS, DDNS, NIS o qualsiasi altro metodo. Ciò include la funzione di bilanciamento del carico DNS disponibile in ONTAP. Quando un database Oracle che utilizza DNFS risolve un nome host in un indirizzo IP, non deve cambiare nelle ricerche successive. Ciò può causare arresti anomali del database Oracle e possibili danni ai dati.

Attivazione di DNFS

Oracle DNFS può funzionare con NFSv3 senza necessità di configurazione oltre all'abilitazione della libreria DNFS (vedere la documentazione di Oracle per il comando specifico richiesto) ma se DNFS non è in grado di stabilire la connettività, può tornare automaticamente al client NFS del kernel. In questo caso, le prestazioni possono essere gravemente compromesse.

Se si desidera utilizzare la multiplazione DNFS su più interfacce, con NFSv4.X, o utilizzare la cifratura, è necessario configurare un file oranfstab. La sintassi è estremamente rigorosa. Piccoli errori nel file possono causare la sospensione dell'avvio o il bypass del file oranfstab.

Al momento della stesura del presente documento, il multipathing DNFS non funziona con NFSv4,1 con le versioni più recenti di Oracle Database. Un file oranfstab che specifica NFSv4,1 come protocollo può utilizzare

solo un'istruzione PATH singola per una determinata esportazione. Il motivo è che ONTAP non supporta il trunking clientID. Le patch dei database Oracle per risolvere questo limite potrebbero essere disponibili in futuro.

L'unico modo per essere certi che DNFS funzioni come previsto è eseguire una query sulle tabelle v$dnfs.

Di seguito è riportato un file oranfstab di esempio che si trova in /etc. Questa è una delle posizioni multiple un file oranfstab può essere posizionato.

```
[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
```

Il primo passo consiste nel verificare che DNFS sia operativo per i filesystem specificati:

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

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

/dbf
NFSv3.0
```

Questo output indica che DNFS è in uso con questi due filesystem, ma **non** significa che oranfstab è operativo. Se fosse presente un errore, DNFS avrebbe scoperto automaticamente i filesystem NFS dell'host e si potrebbe comunque vedere lo stesso output da questo comando.

Il multipathing può essere controllato come segue:

```
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.
```

Di seguito sono riportate le connessioni utilizzate da DNFS. Per ogni voce SVRNAME sono visibili due percorsi e canali. Ciò significa che il multipathing funziona, il che significa che il file oranfstab è stato riconosciuto ed elaborato.

Accesso diretto NFS e file system host

L'utilizzo di DNFS può causare occasionalmente problemi per le applicazioni o le attività degli utenti che si basano sui file system visibili montati sull'host perché il client DNFS accede al file system fuori banda dal sistema operativo host. Il client DNFS può creare, eliminare e modificare i file senza conoscere il sistema operativo.

Quando vengono utilizzate le opzioni di montaggio per i database a istanza singola, consentono la memorizzazione nella cache degli attributi di file e directory, il che significa anche che il contenuto di una directory viene memorizzato nella cache. Pertanto, DNFS può creare un file, e c'è un breve ritardo prima che il sistema operativo rilegga il contenuto della directory e il file diventi visibile all'utente. Questo non è generalmente un problema, ma, in rare occasioni, utility come SAP BR*Tools potrebbero avere problemi. In questo caso, risolvere il problema modificando le opzioni di montaggio in modo da utilizzare le raccomandazioni per Oracle RAC. Questa modifica comporta la disabilitazione di tutta la cache dell'host.

Modificare le opzioni di montaggio solo quando (a) viene utilizzato DNFS e (b) un problema deriva da un ritardo nella visibilità dei file. Se DNFS non è in uso, l'utilizzo delle opzioni di montaggio di Oracle RAC su un database a singola istanza comporta un peggioramento delle prestazioni.

Consultare la nota relativa a `nosharecache` in "Opzioni di montaggio NFS Linux" per un problema DNFS specifico di Linux che può produrre risultati insoliti.

Leasing e blocchi di NFS

NFSv3 è stateless. Ciò significa che il server NFS (ONTAP) non tiene traccia di quali file system sono montati, da chi, o quali blocchi sono realmente presenti.

ONTAP dispone di alcune funzionalità che registreranno i tentativi di mount, quindi si ha un'idea di quali client possono accedere ai dati e potrebbero essere presenti blocchi di avvisi, ma non è garantito che le informazioni siano complete al 100%. Non può essere completo, perché il tracciamento dello stato del client NFS non fa parte dello standard NFSv3.

NFSv4 statefulness

Al contrario, NFSv4 è stateful. Il server NFSv4 tiene traccia di quali client utilizzano i file system, quali file esistono, quali file e/o aree di file sono bloccati, ecc. Ciò significa che è necessaria una comunicazione regolare tra un server NFSv4 per mantenere aggiornati i dati di stato.

Gli stati più importanti gestiti dal server NFS sono NFSv4 Locks e NFSv4 Leasing, e sono molto interconnessi. Dovete capire come ognuno funziona da se stesso e come si relazionano l'uno con l'altro.

NFSv4 serrature

Con NFSv3, i blocchi sono indicativi. Un client NFS può comunque modificare o eliminare un file "bloccato". Un blocco NFSv3 non scade da solo, deve essere rimosso. Questo crea problemi. Ad esempio, se si dispone di un'applicazione in cluster che crea blocchi NFSv3 e uno dei nodi ha esito negativo, come procedere? È possibile codificare l'applicazione sui nodi sopravvissuti per rimuovere i blocchi, ma come si fa a sapere che questo è sicuro? Il nodo "guasto" potrebbe essere operativo, ma non comunica con il resto del cluster?

Con NFSv4, i blocchi hanno una durata limitata. Finché il client che mantiene i blocchi continua il check-in con il server NFSv4, nessun altro client è autorizzato ad acquisire tali blocchi. Se un client non riesce a eseguire il check in con NFSv4, i blocchi vengono revocati dal server e gli altri client potranno richiedere e ottenere i blocchi.

NFSv4 leasing

I blocchi NFSv4 sono associati a un lease NFSv4. Quando un client NFSv4 stabilisce una connessione con un server NFSv4, ottiene un lease. Se il client ottiene un blocco (ci sono molti tipi di blocchi) allora il blocco è associato al lease.

Questo lease ha un timeout definito. Per impostazione predefinita, ONTAP imposta il valore di timeout su 30 secondi:

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

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

Ciò significa che un client NFSv4 deve effettuare il check-in con il server NFSv4 ogni 30 secondi per rinnovare i propri leasing.

Il leasing viene rinnovato automaticamente da qualsiasi attività, quindi se il client sta lavorando non è necessario eseguire operazioni di aggiunta. Se un'applicazione diventa silenziosa e non sta svolgendo un lavoro reale, sarà necessario eseguire una sorta di operazione keep-alive (chiamata SEQUENZA). In sostanza, è solo dire "sono ancora qui, ti prego di rinnovare i miei contratti di leasing".

```
 *Question:* What happens if you lose network connectivity for 31 seconds?
NFSv3 è stateless. Non si aspetta la comunicazione dai clienti. NFSv4 è
stateful, e una volta trascorso il periodo di leasing, il lease scade, i
blocchi vengono revocati e i file bloccati vengono resi disponibili ad
altri client.
```

Con NFSv3, è possibile spostare i cavi di rete, riavviare gli switch di rete, apportare modifiche alla configurazione e assicurarsi che non si verifichi alcun problema. Normalmente, le applicazioni aspettavano solo pazientemente che la connessione di rete funzionasse di nuovo.

Con NFSv4, avete 30 secondi (a meno che non abbiate aumentato il valore di quel parametro all'interno di ONTAP) per completare il vostro lavoro. Se si supera questo limite, il tempo di leasing è scaduto. In genere si verificano arresti anomali delle applicazioni.

Ad esempio, se si dispone di un database Oracle e si verifica una perdita di connettività di rete (talvolta chiamata "partizione di rete") che supera il timeout del lease, il database verrà arrestato.

Di seguito viene riportato un esempio di ciò che accade nel registro degli avvisi di Oracle:

```
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
```

Se si esaminano i syslogs, si dovrebbero vedere alcuni di questi errori:

```
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!
```

I messaggi di registro sono in genere il primo segno di un problema, diverso dal blocco dell'applicazione. In genere, durante l'interruzione della rete non viene visualizzato nulla, poiché i processi e il sistema operativo stesso sono bloccati e tentano di accedere al file system NFS.

Gli errori vengono visualizzati dopo che la rete è nuovamente operativa. Nell'esempio precedente, una volta ristabilita la connettività, il sistema operativo tentava di riacquisire i blocchi, ma era troppo tardi. Il leasing era scaduto e i blocchi sono stati rimossi. Ciò genera un errore che si propaga fino al livello Oracle e causa il messaggio nel registro degli avvisi. È possibile che vengano visualizzate variazioni su questi modelli a seconda della versione e della configurazione del database.

Riassumendo, NFSv3 tollera l'interruzione di rete, ma NFSv4 è più sensibile e impone un periodo di leasing definito.

Cosa succede se un timeout di 30 secondi non è accettabile? Cosa succede se si gestisce una rete a variazione dinamica in cui gli switch vengono riavviati o i cavi vengono ricollocati e il risultato è un'interruzione occasionale della rete? È possibile scegliere di estendere il periodo di leasing, ma se si desidera farlo richiede una spiegazione di NFSv4 periodi di tolleranza.

NFSv4 periodi di grazia

Se un server NFSv3 viene riavviato, è pronto a servire i/o quasi istantaneamente. Non manteneva alcun tipo di stato sui client. Il risultato è che un'operazione di takeover della ONTAP spesso sembra quasi istantanea. Quando un controller è pronto a iniziare a servire i dati, invia un ARP alla rete che segnala la modifica della topologia. I client normalmente rilevano questo quasi istantaneamente e i dati riprendono a fluire.

NFSv4, tuttavia, produrrà una breve pausa. È solo una parte di come funziona NFSv4.

> Le sezioni seguenti sono aggiornate a partire da ONTAP 9.15,1, ma il comportamento lease e lock e le opzioni di ottimizzazione possono cambiare da versione a versione. Se è necessario ottimizzare i timeout di lease/lock NFSv4, consultare il supporto NetApp per le ultime informazioni.

I server NFSv4 devono tenere traccia dei lease, dei blocchi e di chi utilizza i dati. Se un server NFS si riavvia o perde potenza per un momento o viene riavviato durante l'attività di manutenzione, il risultato è il lease/lock e le altre informazioni del client vengono perse. Il server deve individuare quale client utilizza i dati prima di riprendere le operazioni. È qui che entra in gioco il periodo di grazia.

Se all'improvviso si spegne e riaccende il server NFSv4. Quando viene eseguito il backup, i client che tentano di riprendere io riceveranno una risposta che essenzialmente dice: "Ho perso le informazioni di lease/lock. Vuoi registrare nuovamente i blocchi?" Questo è l'inizio del periodo di grazia. Il valore predefinito è 45 secondi su ONTAP:

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

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

Il risultato è che, dopo un riavvio, un controller sospenderà io mentre tutti i client recuperano i loro lease e blocchi. Al termine del periodo di prova, il server riprenderà le operazioni io.

Questo periodo di tolleranza controlla il recupero del leasing durante le modifiche all'interfaccia di rete, ma esiste un secondo periodo di tolleranza che controlla il recupero durante il failover dello storage, `locking.grace_lease_seconds`. Questa è un'opzione a livello di nodo.

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

Ad esempio, se hai spesso bisogno di eseguire failover LIF ed è necessario ridurre il periodo di tolleranza, cambierai `v4-grace-seconds`. Se si desidera migliorare il tempo di ripresa io durante il failover del controller, è necessario modificare `locking.grace_lease_seconds`.

Alterare questi valori solo con cautela e dopo aver compreso appieno i rischi e le conseguenze. Le pause io coinvolte nelle operazioni di failover e migrazione con NFSv4.X non possono essere evitate del tutto. I periodi di blocco, leasing e grazia fanno parte della RFC NFS. Per molti clienti, NFSv3 è preferibile perché i tempi di failover sono più rapidi.

Timeout leasing vs periodi di grazia

Il periodo di tolleranza e il periodo di leasing sono collegati. Come menzionato sopra, il timeout di lease predefinito è di 30 secondi, il che significa che NFSv4 client devono effettuare il check-in con il server almeno ogni 30 secondi o perdere i lease e, a loro volta, i blocchi. Il periodo di tolleranza esiste per consentire a un server NFS di ricostruire i dati di lease/lock e il valore predefinito è 45 secondi. Il periodo di tolleranza deve essere più lungo del periodo di leasing. In questo modo, un ambiente client NFS progettato per rinnovare i lease almeno ogni 30 secondi avrà la possibilità di effettuare il check-in con il server dopo un riavvio. Un periodo di tolleranza di 45 secondi assicura che tutti quei clienti che si aspettano di rinnovare i loro leasing almeno ogni 30 secondi definitivamente hanno l'opportunità di farlo.

Se un timeout di 30 secondi non è accettabile, è possibile scegliere di prolungare il periodo di leasing.

Se si desidera aumentare il timeout del lease a 60 secondi per resistere a un'interruzione di rete di 60 secondi, sarà necessario aumentare anche il periodo di tolleranza. Ciò significa che si verificheranno pause di i/o più lunghe durante il failover del controller.

Normalmente questo non dovrebbe essere un problema. Gli utenti tipici aggiornano i controller ONTAP solo una o due volte all'anno e il failover non pianificato dovuto a guasti hardware è estremamente raro. Inoltre, se aveste una rete in cui un'interruzione di rete di 60 secondi fosse una possibilità preoccupante e aveste bisogno di un timeout del leasing di 60 secondi, probabilmente non vi opporreste a un raro failover del sistema storage con una pausa di 61 secondi. Hai già riconosciuto che la tua rete è in pausa per più di 60 secondi piuttosto frequentemente.

Caching di NFS

La presenza di una delle seguenti opzioni di montaggio causa la disattivazione della cache host:

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

Queste impostazioni possono avere un grave effetto negativo sulla velocità di installazione del software, l'applicazione di patch e le operazioni di backup/ripristino. In alcuni casi, in particolare con le applicazioni in cluster, queste opzioni sono necessarie come conseguenza inevitabile della necessità di garantire la coerenza della cache in tutti i nodi del cluster. In altri casi, i clienti utilizzano erroneamente questi parametri e il risultato è un inutile danno alle prestazioni.

Molti clienti rimuovono temporaneamente queste opzioni di montaggio durante l'installazione o l'applicazione di patch dei file binari. Questa rimozione può essere eseguita in modo sicuro se l'utente verifica che nessun altro processo stia utilizzando attivamente la directory di destinazione durante il processo di installazione o di applicazione delle patch.

Dimensioni trasferimento NFS

Per impostazione predefinita, ONTAP limita le dimensioni i/o NFS a 64K.

L'i/o casuale con la maggior parte delle applicazioni e dei database utilizza blocchi di dimensioni molto inferiori, ben al di sotto del limite massimo di 64K KB. L'i/o a blocchi di grandi dimensioni è solitamente a parallelismo, pertanto anche il massimo di 64K Gbps non costituisce un limite all'ottenimento della massima larghezza di banda.

Ci sono alcuni carichi di lavoro in cui il massimo di 64K crea un limite. In particolare, le operazioni single-threaded, come l'operazione di backup o ripristino o la scansione di un database completa della tabella,

vengono eseguite più velocemente e in modo più efficiente se il database è in grado di eseguire un numero di i/o inferiore ma maggiore. Le dimensioni ottimali per la gestione i/o per ONTAP sono 256K KB.

Le dimensioni massime di trasferimento per una SVM ONTAP possono essere modificate come segue:

```
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::*>
```

Non diminuire mai la dimensione massima di trasferimento consentita su ONTAP al di sotto del valore rsize/wsize dei file system NFS attualmente montati. In alcuni sistemi operativi, ciò può causare blocchi o addirittura danni ai dati. Ad esempio, se i client NFS sono attualmente impostati su un valore rsize/wsize di 65536, la dimensione massima di trasferimento ONTAP potrebbe essere regolata tra 65536 e 1048576 senza alcun effetto perché i client stessi sono limitati. La riduzione della dimensione massima di trasferimento inferiore a 65536 GB può danneggiare la disponibilità o i dati.

NVFAIL

NVFAIL è una funzionalità di ONTAP che garantisce l'integrità in scenari di failover catastrofici.

I database sono vulnerabili al danneggiamento durante gli eventi di failover dello storage perché mantengono grandi cache interne. Se un evento catastrofico richiede l'imposizione di un failover ONTAP o il forzamento dello switchover MetroCluster, a prescindere dallo stato di salute della configurazione complessiva, il risultato viene riconosciuto in precedenza che le modifiche potrebbero essere effettivamente scartate. Il contenuto dell'array di storage torna indietro nel tempo e lo stato della cache del database non riflette più lo stato dei dati su disco. Questa incoerenza provoca il danneggiamento dei dati.

La memorizzazione nella cache può avvenire a livello di applicazione o di server. Ad esempio, una configurazione Oracle Real Application Cluster (RAC) con server attivi sia su un sito primario che su un sito remoto memorizza nella cache i dati all'interno di Oracle SGA. Un'operazione di switchover forzata che comportava la perdita di dati rischierebbe di danneggiare il database poiché i blocchi archiviati nell'SGA potrebbero non corrispondere ai blocchi su disco.

Un uso meno ovvio della memorizzazione nella cache è a livello del file system del sistema operativo. I blocchi di un file system NFS montato possono essere memorizzati nella cache del sistema operativo. In alternativa, un file system in cluster basato su LUN che si trovano nel sito primario può essere montato sui server nel sito remoto e, ancora una volta, i dati possono essere memorizzati nella cache. In queste situazioni, un errore della NVRAM, un takeover forzato o uno switchover forzato possono danneggiare il file system.

ONTAP protegge i database e i sistemi operativi da questo scenario con NVFAIL e le relative impostazioni.

ASM Reclamation Utility (ASMRU)

ONTAP rimuove in modo efficiente i blocchi azzerati scritti su un file o LUN quando la

compressione inline è abilitata. Utility come Oracle ASM Reclamation Utility (ASRU) funzionano scrivendo zero in estensioni ASM non utilizzate.

In questo modo, gli amministratori di database possono recuperare spazio sull'array di storage dopo l'eliminazione dei dati. ONTAP intercetta gli zero e dealloca lo spazio dal LUN. Il processo di recupero dei dati è estremamente rapido, poiché non viene scritto alcun dato all'interno del sistema di storage.

Dal punto di vista del database, il gruppo di dischi ASM contiene zero; la lettura di tali aree del LUN produce un flusso di zero, tuttavia ONTAP non memorizza gli zero sui dischi. Vengono invece apportate semplici modifiche ai metadati che contrassegnano internamente le aree azzerate del LUN come vuote di qualsiasi dato.

Per motivi simili, il test delle performance che implica dati azzerati non è valido, in quanto i blocchi di zero non vengono effettivamente elaborati come scritture all'interno dello storage array.

Quando si utilizza ASRU, assicurarsi che tutte le patch consigliate da Oracle siano installate.

Configurazione dello storage sui sistemi ASA R2

SAN FC

Allineamento delle LUN

L'allineamento delle LUN si riferisce all'ottimizzazione dell'i/o in relazione al layout del file system sottostante.

I sistemi ASA r2 utilizzano la stessa architettura ONTAP di AFF/ FAS , ma con un modello di configurazione semplificato. I sistemi ASA r2 utilizzano le Storage Availability Zone (SAZ) anziché gli aggregati, ma i principi di allineamento rimangono gli stessi perché ONTAP gestisce il layout dei blocchi in modo coerente su tutte le piattaforme. Tuttavia, tieni presente questi punti specifici ASA:

- I sistemi ASA r2 forniscono percorsi simmetrici attivi-attivi per tutte le LUN, eliminando i problemi di asimmetria dei percorsi durante l'allineamento.
- Per impostazione predefinita, le unità di archiviazione (LUN) sono sottoposte a thin provisioning; l'allineamento non modifica questo comportamento.
- La riserva di snapshot e l'eliminazione automatica degli snapshot possono essere configurate durante la creazione della LUN (ONTAP 9.18.1 e versioni successive).

Su un sistema ONTAP, lo storage è organizzato in 4KB unità. Un blocco 8KB di un database o di un file system deve corrispondere esattamente a due blocchi 4KB. Se un errore nella configurazione LUN sposta l'allineamento di 1KB:1 in entrambe le direzioni, ogni blocco 8KB esisterebbe su tre blocchi di storage 4KB diversi invece che due. Questa disposizione causerebbe un aumento della latenza e causerebbe l'esecuzione di ulteriori i/o all'interno del sistema di storage.

L'allineamento influisce anche sulle architetture LVM. Se un volume fisico all'interno di un gruppo di volumi logici viene definito sull'intero dispositivo del disco (non vengono create partizioni), il primo blocco 4KB sul LUN si allinea con il primo blocco 4KB sul sistema di storage. Questo è un allineamento corretto. I problemi si verificano con le partizioni perché spostano la posizione iniziale in cui il sistema operativo utilizza il LUN. Finché l'offset viene spostato in intere unità di 4KB, il LUN viene allineato.

Negli ambienti Linux, creare gruppi di volumi logici sull'intero dispositivo di unità. Quando è necessaria una partizione, controllare l'allineamento eseguendo `fdisk -u` e verificando che l'inizio di ogni partizione sia un

multiplo di otto. Ciò significa che la partizione inizia da un multiplo di otto settori a 512 byte, ovvero 4KB.

Vedere anche la discussione sull'allineamento dei blocchi di compressione nella sezione "Efficienza". Qualsiasi layout allineato ai limiti del blocco di compressione 8KB è allineato ai limiti 4KB.

Avvertenze di disallineamento

La registrazione di ripristino del database/transazioni genera di solito un i/o non allineato che può causare avvisi fuorvianti riguardo ai LUN disallineati su ONTAP.

La registrazione esegue una scrittura sequenziale del file di registro con scritture di dimensioni variabili. Un'operazione di scrittura del registro che non si allinea ai limiti 4KB non causa normalmente problemi di prestazioni poiché l'operazione di scrittura del registro successiva completa il blocco. Il risultato è che ONTAP è in grado di elaborare quasi tutte le scritture come blocchi da 4KB KB completi, anche se i dati in alcuni blocchi da 4KB KB sono stati scritti in due operazioni separate.

Verificare l'allineamento utilizzando utilità come `sio` O `dd` in grado di generare I/O a una dimensione di blocco definita. Le statistiche di allineamento I/O sul sistema di archiviazione possono essere visualizzate con `stats` comando. Vedere "Verifica dell'allineamento di WAFL" per maggiori informazioni.

L'allineamento negli ambienti Solaris è più complicato. Fare riferimento a. "Configurazione host SAN ONTAP" per ulteriori informazioni.

> Negli ambienti Solaris x86, prestare ulteriore attenzione al corretto allineamento poiché la maggior parte delle configurazioni prevede diversi livelli di partizioni. Le sezioni di partizione di Solaris x86 si trovano solitamente in cima a una tabella di partizioni del record di avvio master standard.

Ulteriori buone pratiche:

- Verificare il firmware HBA e le impostazioni del sistema operativo rispetto allo strumento NetApp Interoperability Matrix Tool (IMT).
- Utilizzare le utilità sanlun per confermare l'integrità e l'allineamento del percorso.
- Per Oracle ASM e LVM, assicurarsi che i file di configurazione (/etc/lvm/lvm.conf, /etc/sysconfig/oracleasm) siano impostati correttamente per evitare problemi di allineamento.

Dimensionamento e numero di LUN

La scelta delle dimensioni ottimali e del numero di LUN da utilizzare è un elemento critico per ottenere performance e gestibilità ottimali dei database Oracle.

Una LUN è un oggetto virtualizzato su ONTAP presente su tutte le unità nella Storage Availability Zone (SAZ) di hosting sui sistemi ASA r2. Di conseguenza, le prestazioni della LUN non sono influenzate dalle sue dimensioni, poiché la LUN sfrutta appieno il potenziale prestazionale della SAZ, indipendentemente dalle dimensioni scelte.

Per comodità, i clienti potrebbero desiderare di utilizzare un LUN di particolari dimensioni. Ad esempio, se un database è costruito su un gruppo di dischi LVM o Oracle ASM composto da due LUN da 1TB GB ciascuno, tale gruppo di dischi deve essere aumentato in incrementi di 1TB TB. Potrebbe essere preferibile costruire il gruppo di dischi da otto LUN da 500GB ciascuno in modo che il gruppo di dischi possa essere aumentato con incrementi più piccoli.

La pratica di stabilire una dimensione LUN standard universale è scoraggiata perché ciò può complicare la

gestibilità. Ad esempio, è possibile che una dimensione LUN standard di 100GB TB sia ottimale quando un database o un datastore è compreso nell'intervallo da 1TB a 2TB TB, ma un database o un datastore di 20TB GB richiederebbe 200 LUN. Ciò significa che i tempi di riavvio del server sono più lunghi, che vi sono più oggetti da gestire nelle varie interfacce utente e che prodotti come SnapCenter devono eseguire la ricerca su molti oggetti. Utilizzando un numero inferiore di LUN di dimensioni maggiori è possibile evitare questi problemi.

- Considerazioni ASA r2:*
- La dimensione massima della LUN per ASA r2 è 128 TB, il che consente di utilizzare meno LUN di dimensioni maggiori senza compromettere le prestazioni.
- ASA r2 utilizza le Storage Availability Zone (SAZ) anziché gli aggregati, ma ciò non modifica la logica di dimensionamento delle LUN per i carichi di lavoro Oracle.
- Il thin provisioning è abilitato per impostazione predefinita; il ridimensionamento delle LUN non comporta interruzioni e non richiede di metterle offline.

Numero di LUN

A differenza delle dimensioni delle LUN, il numero di LUN influisce sulle performance. Spesso le prestazioni delle applicazioni dipendono dalla capacità di eseguire i/o paralleli attraverso il livello SCSI. Di conseguenza, due LUN offrono performance migliori rispetto a una singola LUN. Utilizzare un LVM come Veritas VxVM, Linux LVM2 o Oracle ASM è il metodo più semplice per aumentare il parallelismo.

Con ASA r2, i principi per il conteggio LUN rimangono gli stessi di AFF/ FAS perché ONTAP gestisce l'I/O parallelo in modo simile su tutte le piattaforme. Tuttavia, l'architettura SAN-only di ASA r2 e i percorsi simmetrici attivi-attivi garantiscono prestazioni coerenti su tutte le LUN.

I clienti di NetApp hanno in genere ottenuto il minimo beneficio dall'aumento del numero di LUN oltre i sedici, sebbene i test degli ambienti con dischi a stato solido al 100% con i/o casuali molto intensi abbiano dimostrato un ulteriore miglioramento fino a 64 LUN.

NetApp consiglia quanto segue:

In generale, da quattro a sedici LUN sono sufficienti per supportare le esigenze di I/O di qualsiasi carico di lavoro del database Oracle. Meno di quattro LUN potrebbero creare limitazioni nelle prestazioni a causa delle limitazioni nelle implementazioni SCSI host. L'aumento oltre i sedici LUN raramente migliora le prestazioni, tranne in casi estremi (ad esempio carichi di lavoro SSD con I/O casuale molto elevati).

Posizionamento delle LUN

Il posizionamento ottimale delle LUN del database nei sistemi ASA r2 dipende principalmente dal modo in cui verranno utilizzate le varie funzionalità ONTAP .

Nei sistemi ASA r2, le unità di archiviazione (LUN o namespace NVMe) vengono create da un livello di archiviazione semplificato denominato Storage Availability Zone (SAZ), che funge da pool di archiviazione comuni per una coppia HA.

In genere è presente una sola zona di disponibilità dello storage (SAZ) per coppia HA.

Zone di disponibilità di stoccaggio (SAZ)

Nei sistemi ASA r2, i volumi sono ancora presenti, ma vengono creati automaticamente quando vengono create le unità di archiviazione. Le unità di archiviazione (LUN o namespace NVMe) vengono fornite

direttamente all'interno dei volumi creati automaticamente nelle Storage Availability Zone (SAZ). Questa progettazione elimina la necessità di una gestione manuale dei volumi e rende il provisioning più diretto e semplificato per carichi di lavoro a blocchi come i database Oracle.

SAZ e unità di stoccaggio

Le unità di archiviazione correlate (LUN o namespace NVMe) sono solitamente collocate all'interno di un'unica Storage Availability Zone (SAZ). Ad esempio, un database che richiede 10 unità di archiviazione (LUN) in genere avrebbe tutte e 10 le unità posizionate nella stessa SAZ per semplicità e prestazioni.

- Il comportamento predefinito ASA r2 è l'utilizzo di un rapporto 1:1 tra unità di archiviazione e volumi, ovvero un'unità di archiviazione (LUN) per volume.
- In caso di più di una coppia HA nel sistema ASA r2, le unità di archiviazione (LUN) per un dato database possono essere distribuite su più SAZ per ottimizzare l'utilizzo e le prestazioni del controller.

Nel contesto di FC SAN, qui l'unità di archiviazione si riferisce a LUN.

Gruppi di coerenza (CG), LUN e snapshot

In ASA r2, i criteri e le pianificazioni degli snapshot vengono applicati a livello di gruppo di coerenza, ovvero una struttura logica che raggruppa più LUN o namespace NVMe per una protezione coordinata dei dati. Un set di dati composto da 10 LUN richiederebbe un solo criterio di snapshot quando tali LUN fanno parte dello stesso gruppo di coerenza.

I gruppi di coerenza garantiscono operazioni di snapshot atomiche su tutti i LUN inclusi. Ad esempio, un database che risiede su 10 LUN o un ambiente applicativo basato su VMware composto da 10 sistemi operativi diversi possono essere protetti come un singolo oggetto coerente se i LUN sottostanti sono raggruppati nello stesso gruppo di coerenza. Se vengono inseriti in gruppi di coerenza diversi, gli snapshot potrebbero essere perfettamente sincronizzati o meno, anche se programmati contemporaneamente.

In alcuni casi, potrebbe essere necessario suddividere un set correlato di LUN in due gruppi di coerenza diversi a causa dei requisiti di ripristino. Ad esempio, un database potrebbe avere quattro LUN per i file di dati e due LUN per i log. In questo caso, un gruppo di coerenza dei file di dati con 4 LUN e un gruppo di coerenza dei log con 2 LUN potrebbero essere l'opzione migliore. Il motivo è la recuperabilità indipendente: il gruppo di coerenza del file di dati potrebbe essere ripristinato selettivamente a uno stato precedente, il che significa che tutti e quattro i LUN verrebbero riportati allo stato dello snapshot, mentre il gruppo di coerenza del log con i suoi dati critici rimarrebbe inalterato.

CG, LUN e SnapMirror

Le policy e le operazioni SnapMirror vengono eseguite, come le operazioni snapshot, sul gruppo di coerenza e non sulla LUN.

La collocazione congiunta di LUN correlate in un singolo gruppo di coerenza consente di creare una singola relazione SnapMirror e di aggiornare tutti i dati contenuti con un singolo aggiornamento. Come per gli snapshot, anche l'aggiornamento sarà un'operazione atomica. La destinazione SnapMirror avrà sicuramente una replica puntuale dei LUN di origine. Se i LUN fossero distribuiti su più gruppi di coerenza, le repliche potrebbero essere coerenti tra loro o meno.

La replica SnapMirror sui sistemi ASA r2 presenta le seguenti limitazioni:

- La replica sincrona SnapMirror non è supportata.
- La sincronizzazione attiva SnapMirror è supportata solo tra due sistemi ASA r2.
- La replica asincrona SnapMirror è supportata solo tra due sistemi ASA r2.
- La replica asincrona SnapMirror non è supportata tra un sistema ASA r2 e un sistema ASA, AFF o FAS o il cloud.

Scopri di più su "Criteri di replica SnapMirror supportati sui sistemi ASA r2".

CG, LUN e QoS

Sebbene la QoS possa essere applicata selettivamente a singole LUN, solitamente è più semplice impostarla a livello di gruppo di coerenza. Ad esempio, tutti i LUN utilizzati dagli ospiti in un dato server ESX potrebbero essere inseriti in un singolo gruppo di coerenza e quindi potrebbe essere applicata una policy QoS adattiva ONTAP . Il risultato è un limite IOPS per TiB auto-scalabile che si applica a tutte le LUN.

Allo stesso modo, se un database richiedesse 100.000 IOPS e occupasse 10 LUN, sarebbe più semplice impostare un singolo limite di 100.000 IOPS su un singolo gruppo di coerenza piuttosto che impostare 10 limiti individuali di 10.000 IOPS, uno su ogni LUN.

Layout CG multipli

In alcuni casi può essere utile distribuire le LUN su più gruppi di coerenza. Il motivo principale è lo striping del controller. Ad esempio, un sistema di archiviazione HA ASA r2 potrebbe ospitare un singolo database Oracle in cui è richiesta la piena capacità di elaborazione e memorizzazione nella cache di ciascun controller. In questo caso, una progettazione tipica sarebbe quella di posizionare metà delle LUN in un singolo gruppo di coerenza sul controller 1 e l'altra metà delle LUN in un singolo gruppo di coerenza sul controller 2.

Allo stesso modo, per gli ambienti che ospitano numerosi database, la distribuzione delle LUN su più gruppi di coerenza può garantire un utilizzo equilibrato del controller. Ad esempio, un sistema HA che ospita 100 database da 10 LUN ciascuno potrebbe assegnare 5 LUN a un gruppo di coerenza sul controller 1 e 5 LUN a un gruppo di coerenza sul controller 2 per database. Ciò garantisce un caricamento simmetrico man mano che vengono forniti database aggiuntivi.

Nessuno di questi esempi, però, prevede un rapporto LUN-gruppo di coerenza pari a 1:1. L'obiettivo rimane quello di ottimizzare la gestibilità raggruppando logicamente le LUN correlate in gruppi di coerenza.

Un esempio in cui un rapporto LUN/gruppo di coerenza di 1:1 ha senso sono i carichi di lavoro containerizzati, in cui ogni LUN potrebbe rappresentare in realtà un singolo carico di lavoro che richiede policy di snapshot e replica separate e quindi deve essere gestito individualmente. In questi casi, un rapporto 1:1 potrebbe essere ottimale.

Ridimensionamento LUN e ridimensionamento LVM

Quando un file system basato su SAN o un gruppo di dischi Oracle ASM raggiunge il limite di capacità su ASA r2, sono disponibili due opzioni per aumentare lo spazio disponibile:

- Aumentare le dimensioni delle LUN (unità di archiviazione) esistenti
- Aggiungere un nuovo LUN a un gruppo di dischi ASM o a un gruppo di volumi LVM esistente e aumentare il volume logico contenuto

Sebbene il ridimensionamento LUN sia supportato su ASA r2, in genere è preferibile utilizzare un Logical Volume Manager (LVM) come Oracle ASM. Uno dei motivi principali per cui esistono gli LVM è quello di evitare la necessità di frequenti ridimensionamenti delle LUN. Con un LVM, più LUN vengono combinati in un pool virtuale di storage. I volumi logici ricavati da questo pool possono essere facilmente ridimensionati senza influire sulla configurazione di archiviazione sottostante.

Ulteriori vantaggi derivanti dall'utilizzo di LVM o ASM includono:

- Ottimizzazione delle prestazioni: distribuisce l'I/O su più LUN, riducendo gli hotspot.
- Flessibilità: aggiungi nuove LUN senza interrompere i carichi di lavoro esistenti.
- Migrazione trasparente: ASM o LVM possono spostare le estensioni su nuove LUN per il bilanciamento o la suddivisione in livelli senza tempi di inattività dell'host.

Considerazioni chiave su ASA r2:

- Il ridimensionamento LUN viene eseguito a livello di unità di archiviazione all'interno di una Storage VM (SVM) utilizzando la capacità della Storage Availability Zone (SAZ).
- Per Oracle, la procedura consigliata è quella di aggiungere LUN ai gruppi di dischi ASM anziché ridimensionare le LUN esistenti, per mantenere lo striping e il parallelismo.

Striping LVM

Lo striping LVM si riferisce alla distribuzione dei dati su più LUN. Il risultato è un significativo miglioramento delle performance per molti database.

Prima dell'era dei dischi flash, era stato utilizzato lo striping per superare i limiti di performance dei dischi rotanti. Ad esempio, se un sistema operativo deve eseguire un'operazione di lettura a 1MB bit, la lettura di 1MB GB di dati da un'unica unità richiederebbe un'ampia ricerca e lettura della testina dell'unità poiché il sistema 1MB viene trasferito lentamente. Se quei 1MB TB di dati sono stati suddivisi in 8 LUN, il sistema operativo potrebbe emettere otto operazioni di lettura 128K in parallelo, riducendo il tempo necessario per completare il trasferimento da 1MB GB.

Lo striping con unità rotanti era più difficile perché era necessario conoscere in anticipo il modello I/O. Se lo striping non fosse stato regolato correttamente per i veri modelli I/O, le configurazioni con striping potrebbero compromettere le prestazioni. Con i database Oracle, e in particolare con le configurazioni di storage all-flash, lo striping è molto più semplice da configurare e ha dimostrato di migliorare notevolmente le prestazioni.

Per impostazione predefinita, i gestori di volume logici, come lo stripe di Oracle ASM, ma il sistema operativo LVM nativo non lo fanno. Alcune di esse collegano più LUN insieme come un dispositivo concatenato, il che comporta file di dati che esistono su un solo dispositivo LUN. Ciò causa punti caldi. Le altre implementazioni LVM sono impostate per impostazione predefinita su estensioni distribuite. Questo è simile allo striping, ma è più grossolano. I LUN nel gruppo di volumi vengono suddivisi in porzioni di grandi dimensioni, chiamate estensioni e generalmente misurati in molti megabyte, e i volumi logici vengono quindi distribuiti tra tali estensioni. Il risultato è un i/o casuale per un file dovrebbe essere ben distribuito tra i LUN, ma le operazioni i/o sequenziali non sono così efficienti come potrebbero essere.

L'i/o delle applicazioni che richiedono elevate performance è quasi sempre (a) in unità delle dimensioni dei blocchi di base o (b) un megabyte.

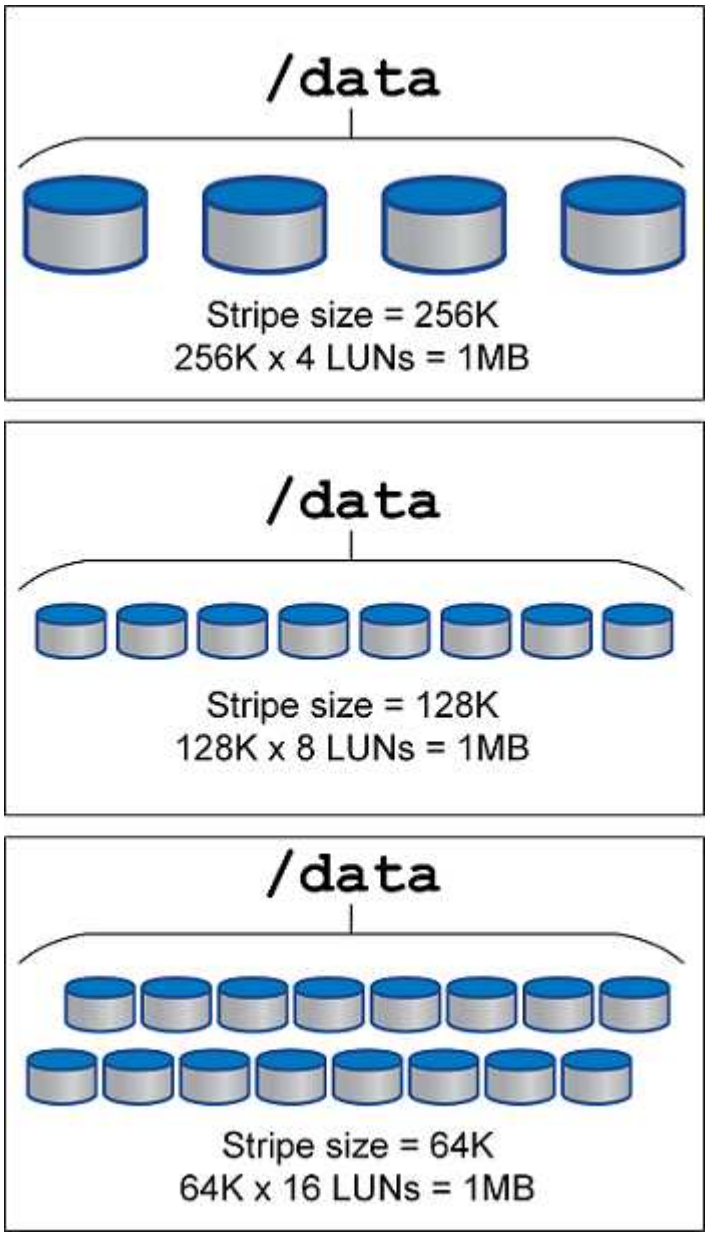
L'obiettivo principale di una configurazione con striping è quello di garantire che l'i/o a file singolo possa essere eseguito come una singola unità, mentre l'i/o a blocchi multipli, di dimensioni pari a 1MB GB, può essere parallelizzato in modo uniforme tra tutti i LUN del volume con striping. Ciò significa che la dimensione dello

stripe non deve essere inferiore alla dimensione del blocco del database e che la dimensione dello stripe moltiplicata per il numero di LUN deve essere 1MB.

Best practice per lo striping LVM con il database Oracle:

- Dimensione stripe ≥ dimensione blocco database.
- Dimensione stripe * numero di LUN ≈ 1 MB per un parallelismo ottimale.
- Utilizzare più LUN per gruppo di dischi ASM per massimizzare la produttività ed evitare hotspot.

La figura seguente mostra tre possibili opzioni per la regolazione delle dimensioni dello stripe e della larghezza. Il numero di LUN viene selezionato per soddisfare i requisiti di prestazioni come descritto sopra, ma in tutti i casi i dati totali all'interno di uno stripe singolo sono 1MB.



NVFAIL

NVFAIL è una funzionalità ONTAP che garantisce l'integrità dei dati durante scenari di failover catastrofici.

Questa funzionalità è ancora applicabile sui sistemi ASA r2, anche se ASA r2 utilizza un'architettura SAN semplificata (SAZ e unità di archiviazione anziché volumi).

I database sono vulnerabili al danneggiamento durante gli eventi di failover dell'archiviazione perché mantengono grandi cache interne. Se un evento catastrofico richiede di forzare un failover ONTAP , indipendentemente dallo stato di salute della configurazione complessiva, il risultato è che le modifiche precedentemente riconosciute possono essere effettivamente ignorate. Il contenuto dell'array di archiviazione torna indietro nel tempo e lo stato della cache del database non riflette più lo stato dei dati sul disco. Questa incoerenza provoca la corruzione dei dati.

La memorizzazione nella cache può avvenire a livello di applicazione o di server. Ad esempio, una configurazione Oracle Real Application Cluster (RAC) con server attivi sia su un sito primario che su uno remoto memorizza nella cache i dati all'interno di Oracle SGA. Un'operazione di failover forzato che comportasse la perdita di dati esporrebbe il database al rischio di danneggiamento, poiché i blocchi memorizzati nell'SGA potrebbero non corrispondere ai blocchi sul disco.

Un utilizzo meno ovvio della memorizzazione nella cache è a livello del file system del sistema operativo. Un file system in cluster basato su LUN ubicati nel sito primario potrebbe essere montato sui server nel sito remoto e, ancora una volta, i dati potrebbero essere memorizzati nella cache. In queste situazioni, un guasto della NVRAM o un'acquisizione forzata potrebbero causare il danneggiamento del file system.

ONTAP protegge i database e i sistemi operativi da questo scenario utilizzando NVFAIL e le impostazioni associate, che segnalano all'host di invalidare i dati memorizzati nella cache e di rimontare i file system interessati dopo il failover. Questo meccanismo si applica ai LUN e agli spazi dei nomi ASA r2 proprio come avviene su AFF/ FAS.

Considerazioni chiave su ASA r2:

- NVFAIL opera a livello LUN (unità di archiviazione), non a livello SAZ.
- Per i database Oracle, NVFAIL deve essere abilitato su tutte le LUN che ospitano componenti critici (file di dati, redo log, file di controllo).
- MetroCluster non è supportato su ASA r2, quindi NVFAIL si applica principalmente agli scenari di failover HA locale.
- NFS non è supportato su ASA r2, pertanto le considerazioni su NVFAIL si applicano solo ai carichi di lavoro basati su SAN (FC/iSCSI/NVMe).

Utilità di recupero ASM (ASRU)

ONTAP su ASA r2 rimuove in modo efficiente i blocchi azzerati scritti su una LUN (unità di archiviazione) quando è abilitata la compressione in linea. Utilità come Oracle ASM Reclamation Utility (ASRU) funzionano scrivendo zeri nelle estensioni ASM non utilizzate.

Ciò consente agli amministratori di database di recuperare spazio sull'array di archiviazione dopo l'eliminazione dei dati. ONTAP intercetta gli zeri e dealloca lo spazio dalla LUN. Il processo di recupero è estremamente rapido perché nel sistema di archiviazione non vengono scritti dati effettivi.

Dal punto di vista del database, il gruppo di dischi ASM contiene zero; la lettura di tali aree del LUN produce

un flusso di zero, tuttavia ONTAP non memorizza gli zero sui dischi. Vengono invece apportate semplici modifiche ai metadati che contrassegnano internamente le aree azzerate del LUN come vuote di qualsiasi dato.

Per motivi simili, il test delle performance che implica dati azzerati non è valido, in quanto i blocchi di zero non vengono effettivamente elaborati come scritture all'interno dello storage array.

Considerazioni chiave ASRU con ASA r2 ONTAP:

- Funziona allo stesso modo di AFF/ FAS per i carichi di lavoro SAN perché ASA r2 è solo a blocchi.
- Si applica ai LUN e agli spazi dei nomi NVMe forniti all'interno delle SAZ.
- Non esistono volumi FlexVol , ma il comportamento di recupero del blocco zero è identico.

Quando si utilizza ASRU, assicurarsi che tutte le patch consigliate da Oracle siano installate.

Virtualizzazione

La virtualizzazione dei database con VMware, Oracle OLVM o KVM è una scelta sempre più comune per i clienti NetApp che hanno scelto la virtualizzazione anche per i database mission-critical.

Supportabilità

Esistono numerosi preconcetti sui criteri di supporto Oracle per la virtualizzazione, in particolare per i prodotti VMware. Non è raro che Oracle Outright non supporti la virtualizzazione. Questa nozione non è corretta e porta alla perdita di opportunità per trarre vantaggio dalla virtualizzazione. Oracle Doc ID 249212,1 illustra i requisiti effettivi e raramente viene considerato un problema da parte dei clienti.

Se si verifica un problema su un server virtualizzato e il supporto di Oracle non lo conosce, al cliente potrebbe essere richiesto di riprodurre il problema sull'hardware fisico. Un cliente Oracle che utilizza una versione all'avanguardia di un prodotto potrebbe non voler utilizzare la virtualizzazione a causa di potenziali problemi di supportabilità, ma questa situazione non è stata un problema reale per i clienti che utilizzano versioni di prodotti Oracle generalmente disponibili.

Presentazione storage

I clienti che stanno considerando la virtualizzazione dei propri database devono basare le proprie decisioni di storage sulle esigenze aziendali. Sebbene questa affermazione sia generalmente vera per tutte le decisioni IT, è particolarmente importante per i progetti di database, poiché le dimensioni e l'ambito dei requisiti variano notevolmente.

Sono disponibili tre opzioni di base per la presentazione dello storage:

- LUN virtualizzate nei datastore di hypervisor
- LUN iSCSI gestite dall'iniziatore iSCSI sulla macchina virtuale, non dall'hypervisor
- File system NFS montati dalla macchina virtuale (non da un datastore basato su NFS)
- Mappatura diretta dei dispositivi. Gli RDM VMware sono svantaggiati dai clienti, ma spesso i dispositivi fisici sono ancora mappati in modo simile direttamente con la virtualizzazione KVM e OLVM.

Performance

Il metodo di presentazione dello storage a un guest virtualizzato non influisce in genere sulle prestazioni. I sistemi operativi host, i driver di rete virtualizzati e le implementazioni del datastore degli hypervisor sono tutti altamente ottimizzati e generalmente possono consumare tutta la larghezza di banda della rete FC o IP disponibile tra l'hypervisor e il sistema storage, purché vengano seguite le Best practice di base. In alcuni casi, ottenere prestazioni ottimali può essere leggermente più semplice utilizzando un approccio di presentazione dello storage rispetto a un altro, ma il risultato finale dovrebbe essere comparabile.

Gestibilità

Il fattore chiave nella scelta di come presentare lo storage a un guest virtualizzato è la manovrabilità. Non esiste un metodo giusto o sbagliato. L'approccio migliore dipende dalle esigenze operative, dalle competenze e dalle preferenze DELL'IT.

I fattori da prendere in considerazione includono:

- **Trasparenza.** quando una VM gestisce i propri file system, è più facile per un amministratore di database o un amministratore di sistema identificare l'origine dei file system per i propri dati. L'accesso ai file system e ai LUN non avviene in modo diverso rispetto a un server fisico.
- **Coerenza.** quando una VM è proprietaria dei file system, l'utilizzo o il mancato utilizzo di un livello di hypervisor influisce sulla gestibilità. È possibile utilizzare le stesse procedure per il provisioning, il monitoraggio, la protezione dei dati e così via nell'intero ambiente, inclusi ambienti virtualizzati e non.

  D'altra parte, in un data center altrimenti virtualizzato al 100% potrebbe essere preferibile utilizzare anche lo storage basato su datastore per l'intera impronta, sulla stessa logica sopra menzionata, la coerenza, la capacità di utilizzare le stesse procedure per il provisioning, la protezione, il montoring e la protezione dei dati.

- **Stabilità e risoluzione dei problemi.** quando una VM possiede i propri file system, fornire prestazioni buone e stabili e risolvere i problemi è più semplice perché l'intero stack di storage è presente sulla VM. L'unico ruolo dell'hypervisor è il trasporto di frame FC o IP. Quando un datastore è incluso in una configurazione, complica la configurazione introducendo un altro insieme di timeout, parametri, file di log e potenziali bug.
- **Portabilità.** quando una VM è proprietaria dei suoi file system, il processo di spostamento di un ambiente Oracle diventa molto più semplice. I file system possono essere spostati facilmente tra guest virtualizzati e non.
- **Vendor lock-in.** una volta posizionati i dati in un datastore, diventa difficile utilizzare un hypervisor diverso o estrarre i dati dall'ambiente virtualizzato.
- **Abilitazione snapshot.** le procedure di backup tradizionali in un ambiente virtualizzato possono diventare un problema a causa della larghezza di banda relativamente limitata. Ad esempio, un trunk 10GbE a quattro porte potrebbe essere sufficiente per supportare le esigenze quotidiane di prestazioni di molti database virtualizzati, ma tale trunk non sarebbe sufficiente per eseguire backup utilizzando RMAN o altri prodotti di backup che richiedono lo streaming di una copia di dimensioni complete dei dati. Il risultato è che un ambiente virtualizzato sempre più consolidato ha bisogno di eseguire backup tramite snapshot di storage. In questo modo si evita la necessità di sovrascrivere la configurazione dell'hypervisor solo per supportare i requisiti di larghezza di banda e CPU nella finestra di backup.

  L'utilizzo di file system guest facilita a volte l'utilizzo di backup e ripristini basati su snapshot, poiché gli oggetti storage che necessitano di protezione possono essere indirizzati più facilmente. Tuttavia, esiste un numero sempre maggiore di prodotti di data Protection di virtualizzazione che si integrano perfettamente con datastore e snapshot. La strategia di backup deve essere considerata attentamente prima di prendere una decisione su come presentare lo storage a un host virtualizzato.

Driver paravirtualizzati

Per prestazioni ottimali, l'uso di driver di rete paravirtualizzati è fondamentale. Quando si utilizza un datastore, è necessario un driver SCSI paravirtualizzato. Un driver di dispositivo paravirtualizzato consente a un guest di integrarsi più profondamente nell'hypervisor, invece di un driver emulato in cui l'hypervisor spende più tempo CPU che imita il comportamento dell'hardware fisico.

Overcommit RAM

L'overcommit della RAM implica la configurazione di una quantità di RAM virtualizzata su vari host superiore a quella presente sull'hardware fisico. In caso contrario, si potrebbero verificare problemi di prestazioni imprevisti. Quando si virtualizza un database, i blocchi sottostanti di Oracle SGA non devono essere sostituiti con lo storage dall'hypervisor. Ciò causa risultati di prestazioni altamente instabili.

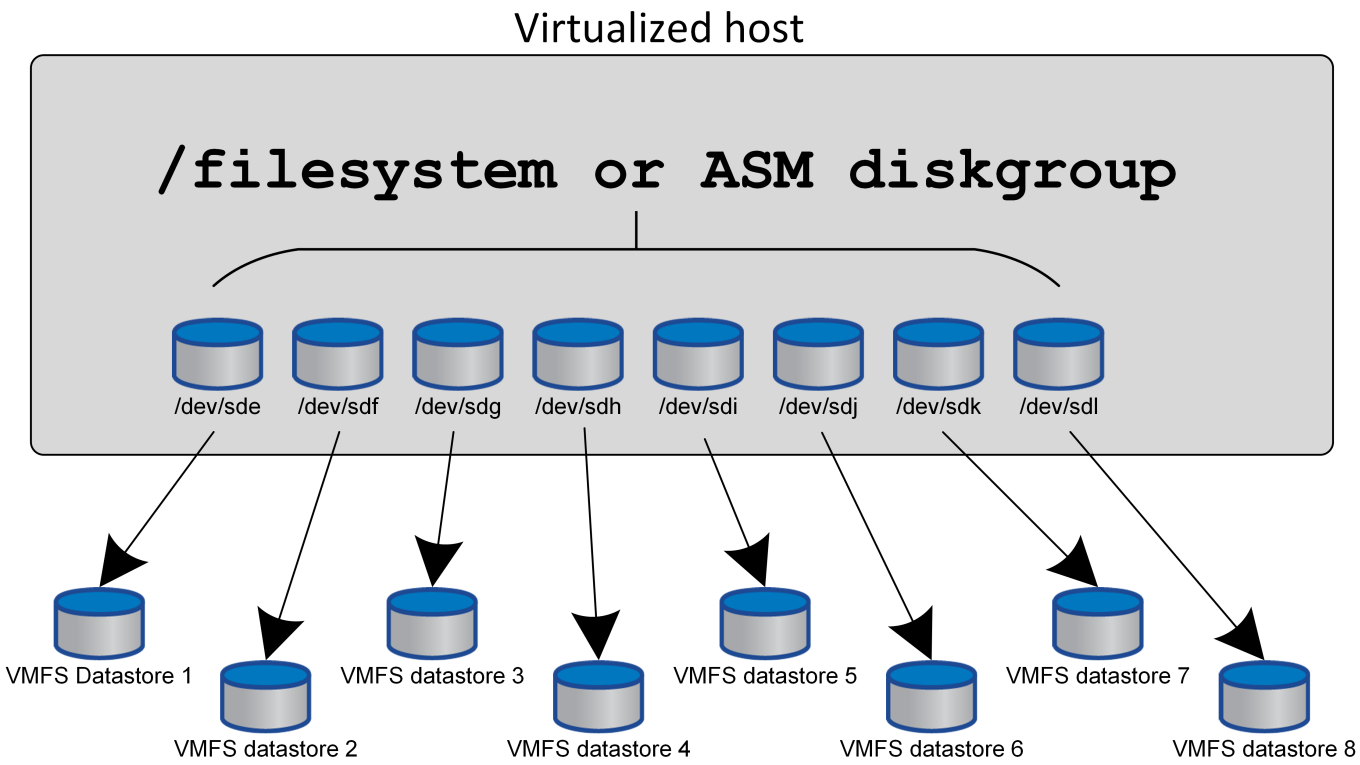
Striping dei datastore

Quando si utilizzano database con datastore, c'è un fattore critico da considerare in relazione allo striping delle performance.

Le tecnologie dei datastore come VMFS sono in grado di estendersi su più LUN, ma non su dispositivi suddivisi in striping. I LUN sono concatenati. Il risultato finale può essere costituito da hot spot LUN. Ad esempio, un database Oracle tipico potrebbe avere un gruppo di dischi ASM di 8 LUN. È possibile eseguire il provisioning di tutte le 8 LUN virtualizzate in un datastore VMFS da 8 LUN, senza tuttavia alcuna garanzia su quali LUN risiedono i dati. La configurazione risultante potrebbe essere tutta una LUN virtualizzata da 8 GB che occupa una singola LUN nel datastore VMFS. Ciò si traduce in un collo di bottiglia per le prestazioni.

Lo striping è in genere necessario. Con alcuni hypervisor, incluso KVM, è possibile creare un datastore utilizzando lo striping LVM come descritto "qui". Con VMware, l'architettura appare un po' diversa. Ogni LUN virtualizzata deve essere posizionata in un datastore VMFS diverso.

Ad esempio:



Il driver principale di questo approccio non è ONTAP, ma è dovuto alla limitazione intrinseca del numero di operazioni che una singola VM o LUN dell'hypervisor può eseguire in parallelo. In genere, un singolo LUN ONTAP può supportare un maggior numero di IOPS rispetto a quello richiesto da un host. Il limite di prestazioni di un singolo LUN è quasi universalmente il risultato del sistema operativo host. Il risultato è che per soddisfare le esigenze di performance della maggior parte dei database sono necessarie LUN comprese tra 4 e 8 GB.

Le architetture VMware devono pianificare con attenzione le proprie architetture per garantire che questo approccio non soddisfi i massimi di datastore e/o percorso LUN. Inoltre, non è necessario un set univoco di datastore VMFS per ogni database. L'esigenza principale consiste nel garantire che ogni host disponga di un set pulito di percorsi io da 4-8 GB dalle LUN virtualizzate alle LUN di backend sul sistema storage stesso. In rare occasioni, anche un numero maggiore di datatores può rivelarsi vantaggioso per richieste di performance veramente estreme, ma le LUN da 4-8 GB sono in genere sufficienti per il 95% di tutti i database. Un singolo volume ONTAP contenente 8 LUN può supportare fino a 250.000 IOPS casuali con blocchi Oracle con una tipica configurazione di sistema operativo/ONTAP/rete.

Tiering

Panoramica

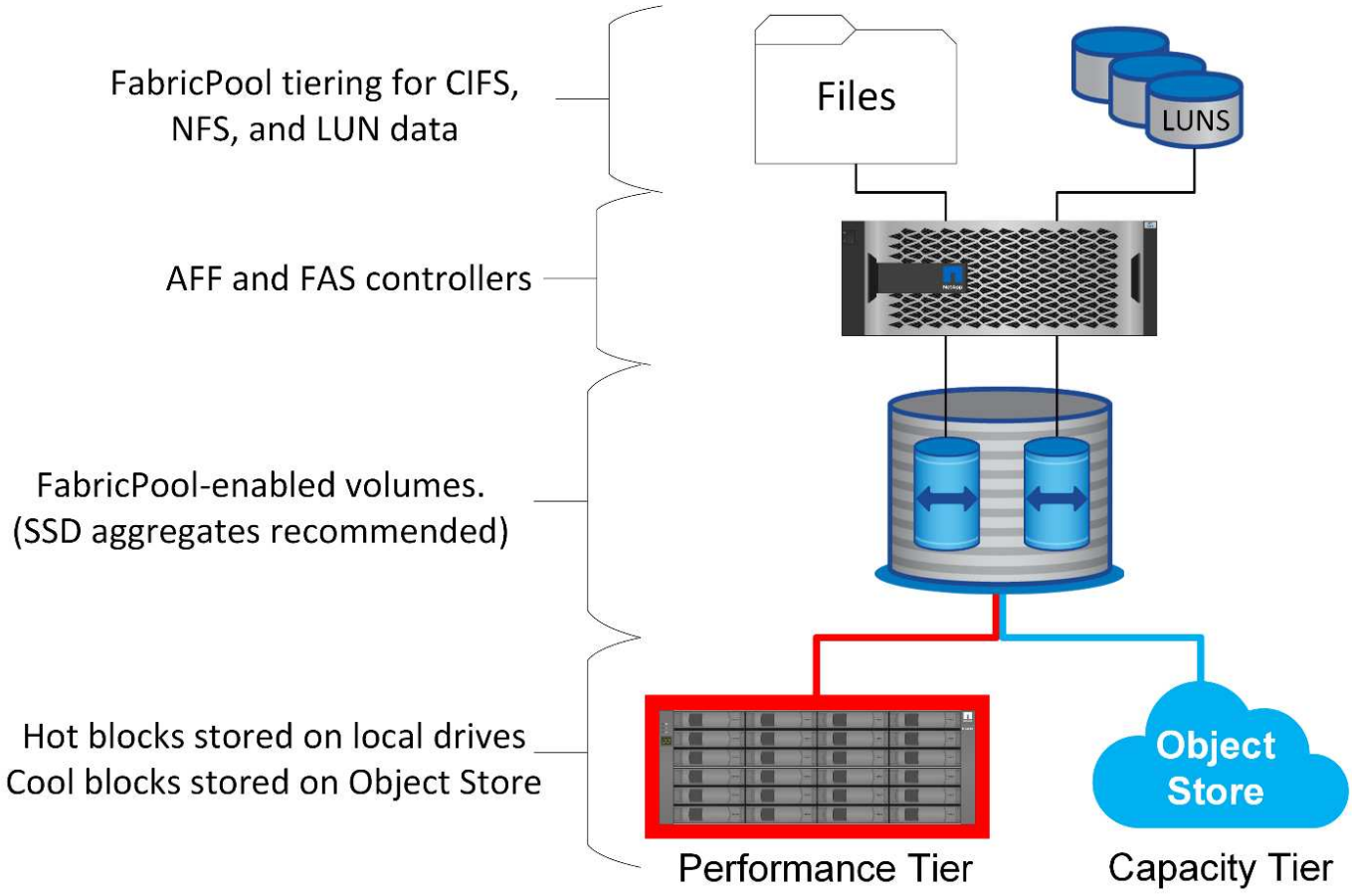
La comprensione dell'impatto del tiering FabricPool su Oracle e altri database richiede una conoscenza dell'architettura FabricPool di basso livello.

Architettura

FabricPool è una tecnologia di tiering che classifica i blocchi come "hot" o "cool" e li colloca nel Tier di storage più appropriato. Il Tier di performance è nella maggior parte dei casi collocato nello storage SSD e ospita i blocchi di dati "hot". Il Tier di capacità si trova in un archivio di oggetti e ospita i blocchi di dati "cool". Il supporto per lo storage a oggetti include NetApp StorageGRID, ONTAP S3, archiviazione BLOB di Microsoft

Azure, il servizio di storage a oggetti Alibaba Cloud, archiviazione a oggetti IBM Cloud, archiviazione Google Cloud e Amazon AWS S3.

Sono disponibili più policy di tiering che controllano le modalità di classificazione dei blocchi come "hot" o "cool", che possono essere impostate in base al volume e modificate secondo necessità. Solo i blocchi di dati vengono spostati tra i Tier di performance e capacità. I metadati che definiscono la struttura LUN e del file system rimangono sempre sul Tier di performance. Di conseguenza, la gestione è centralizzata su ONTAP. I file e le LUN non appaiono diversi dai dati memorizzati in qualsiasi altra configurazione ONTAP. Il controller NetApp AFF o FAS applica le policy definite per spostare i dati nel Tier appropriato.



Provider di archivi di oggetti

I protocolli di storage a oggetti utilizzano semplici richieste HTTP o HTTPS per la memorizzazione di un grande numero di oggetti dati. L'accesso allo storage a oggetti deve essere affidabile, poiché l'accesso ai dati da parte di ONTAP dipende dalla puntuale manutenzione delle richieste. Le opzioni includono le opzioni Amazon S3 Standard e accesso poco frequente, Microsoft Azure Hot e Cool Blob Storage, IBM Cloud e Google Cloud. Le opzioni di archiviazione come Amazon Glacier e Amazon Archive non sono supportate, perché il tempo necessario per recuperare i dati può superare le tolleranze dei sistemi operativi e delle applicazioni host.

NetApp StorageGRID è anche supportato e rappresenta una soluzione di livello Enterprise ottimale. Si tratta di un sistema storage a oggetti dalle performance elevate, scalabile e altamente sicuro, in grado di fornire ridondanza geografica per i dati FabricPool nonché per altre applicazioni di archivi di oggetti che hanno sempre più probabilità di far parte di ambienti applicativi Enterprise.

StorageGRID può anche ridurre i costi evitando le spese di uscita imposte da molti provider di cloud pubblici per la lettura dei dati di nuovo dai propri servizi.

Dati e metadati

Si noti che il termine "dati" si applica in questo caso ai blocchi di dati effettivi, non ai metadati. Viene eseguito il tiering solo dei blocchi di dati, mentre i metadati rimangono nel Tier di performance. Inoltre, lo stato di un blocco come caldo o freddo è influenzato solo dalla lettura del blocco di dati effettivo. La semplice lettura del nome, dell'indicatore data e ora o dei metadati di proprietà di un file non influisce sulla posizione dei blocchi di dati sottostanti.

Backup

Anche se FabricPool può ridurre significativamente l'impatto dello storage, non rappresenta di per sé una soluzione di backup. I metadati NetApp WAFL rimangono sempre nel Tier di performance. Se un disastro catastrofico distrugge il Tier di performance, non è possibile creare un nuovo ambiente utilizzando i dati sul Tier di capacità perché non contiene metadati WAFL.

FabricPool, tuttavia, può entrare a far parte di una strategia di backup. Ad esempio, FabricPool può essere configurato con la tecnologia di replica NetApp SnapMirror. Ciascuna metà del mirror può avere la propria connessione a una destinazione dello storage a oggetti. Il risultato sono due copie indipendenti dei dati. La copia primaria è costituita dai blocchi sul Tier di performance e dai blocchi associati nel Tier di capacità, mentre la replica è un secondo set di blocchi di performance e capacità.

Policy di tiering

Policy di tiering

In ONTAP sono disponibili quattro criteri che controllano il modo in cui i dati Oracle sul livello di prestazioni diventano candidati per il trasferimento al livello di capacità.

Solo snapshot

Il `snapshot-only` `tiering-policy` si applica solo ai blocchi non condivisi con il file system attivo. Essenzialmente si traduce in tiering dei backup del database. I blocchi diventano candidati per il tiering dopo la creazione di uno snapshot e il blocco viene quindi sovrascritto, generando un blocco presente solo all'interno dello snapshot. Il ritardo prima di un `snapshot-only` il blocco è considerato freddo e controllato da `tiering-minimum-cooling-days` impostazione del volume. L'intervallo a partire da ONTAP 9,8 è compreso tra 2 e 183 giorni.

Molti set di dati hanno tassi di cambiamento bassi, con conseguenti risparmi minimi derivanti da questa policy. Ad esempio, un database tipico osservato con ONTAP ha un tasso di variazione inferiore al 5% alla settimana. I log di archivio dei database possono occupare spazio esteso, ma in genere continuano a esistere nel file system attivo e pertanto non possono essere candidati per il tiering in base a questa policy.

Automatico

Il `auto` la policy di tiering estende il tiering sia a blocchi specifici di snapshot che a blocchi nel file system attivo. Il ritardo prima che un blocco venga considerato freddo è controllato dall' `tiering-minimum-cooling-days` impostazione del volume. L'intervallo a partire da ONTAP 9,8 è compreso tra 2 e 183 giorni.

Questo approccio abilita opzioni di tiering che non sono disponibili con `snapshot-only` policy. Ad esempio, un criterio di protezione dei dati potrebbe richiedere la conservazione di 90 giorni di determinati file di registro. L'impostazione di un periodo di raffreddamento di 3 giorni comporta il tiering di tutti i file di registro precedenti a 3 giorni dal livello delle prestazioni. Questa azione libera spazio sostanziale sul Tier delle performance, consentendoti comunque di visualizzare e gestire tutti e 90 i giorni di dati.

Nessuno

Il `none` la policy di tiering impedisce il tiering di blocchi aggiuntivi dal layer di storage, ma i dati ancora presenti nel tier di capacità rimangono nel tier di capacità fino a quando non vengono letti. Se quindi il blocco viene letto, viene tirato indietro e posizionato nel Tier di performance.

Il motivo principale per cui si utilizza `none` la policy di tiering impedisce il tiering dei blocchi, ma nel tempo potrebbe risultare utile modificarli. Ad esempio, supponiamo che un set di dati specifico venga suddiviso in Tier per il livello di capacità, ma sorge un'esigenza inaspettata di funzionalità di performance complete. La policy può essere modificata per impedire qualsiasi tiering aggiuntivo e per confermare che i blocchi letti nuovamente quando l'io aumenta rimangono nel Tier di performance.

Tutto

Il `all` la policy di tiering sostituisce `backup` Policy in data ONTAP 9,6. Il `backup` Policy applicata solo ai volumi di data Protection, che significa destinazione SnapMirror o NetApp SnapVault. Il `all` le funzioni dei criteri sono identiche, ma non si limitano ai volumi di protezione dei dati.

Grazie a questa policy, i blocchi vengono immediatamente considerati COOL e possono essere immediatamente suddivisi in Tier nel livello di capacità.

Questo criterio è particolarmente appropriato per i backup a lungo termine. Può anche essere utilizzato come forma di gestione gerarchica dello storage (HSM, Hierarchical Storage Management). In passato, HSM veniva comunemente utilizzato per eseguire il tiering dei blocchi di dati di un file su nastro, mantenendo il file stesso visibile nel file system. Un volume FabricPool con `all` il criterio consente di archiviare i file in un archivio visibile e gestibile pur non occupando quasi nessuno spazio nel livello di storage locale.

Criteri di recupero

I criteri di tiering controllano i blocchi di database Oracle sottoposti a tiering dal Tier di performance al Tier di capacità. I criteri di recupero controllano ciò che accade quando viene letto un blocco a cui è stato eseguito il tiering.

Predefinito

Tutti i volumi FabricPool sono inizialmente impostati su `default`, il che significa che il comportamento è controllato da `cloud-retrieval-policy. `il comportamento esatto dipende dal criterio di tiering utilizzato.

- `auto`consente di recuperare solo dati letti in modo casuale
- `snapshot-only`- consente di recuperare tutti i dati letti in modo sequenziale o casuale
- `none`- consente di recuperare tutti i dati letti in modo sequenziale o casuale
- `all`- non recuperare i dati dal tier di capacità

A lettura

Impostazione `cloud-retrieval-policy` in lettura sovrascrive il comportamento predefinito, in modo che una lettura di dati a livelli determini il ritorno dei dati al livello di prestazioni.

Ad esempio, un volume potrebbe essere stato leggermente utilizzato per un lungo periodo sotto il `auto` la policy di tiering e la maggior parte dei blocchi ora vengono suddivisi in livelli.

Se una modifica imprevista delle esigenze aziendali richiedeva la scansione ripetuta di alcuni dati per

preparare un determinato rapporto, potrebbe essere opportuno modificare `cloud-retrieval-policy` a. `on-read` per garantire che tutti i dati letti vengano restituiti al livello delle prestazioni, inclusi i dati letti in modo sequenziale e casuale. In questo modo si migliorano le prestazioni dell'i/o sequenziale rispetto al volume.

Promuovi

Il comportamento della policy di promozione dipende dalla policy di tiering. Se la policy di tiering è `auto`, quindi impostare `cloud-retrieval-policy `to `promote` riporta tutti i blocchi dal tier di capacità nella successiva scansione del tiering.

Se la policy di tiering è `snapshot-only`, gli unici blocchi restituiti sono i blocchi associati al file system attivo. Normalmente questo non avrebbe alcun effetto perché gli unici blocchi suddivisi in livelli sotto `snapshot-only` la policy dovrebbe essere costituita da blocchi associati esclusivamente agli snapshot. Nel file system attivo non sono presenti blocchi a livelli.

Se, tuttavia, i dati di un volume sono stati ripristinati da un'operazione SnapRestore di volume o di file-clone da una snapshot, alcuni dei blocchi suddivisi in Tier perché associati solo a snapshot potrebbero ora essere richiesti dal file system attivo. Potrebbe essere opportuno modificare temporaneamente `cloud-retrieval-policy` policy to `promote` per recuperare rapidamente tutti i blocchi richiesti localmente.

Mai

Non recuperare i blocchi dal Tier di capacità.

Strategie di tiering

Tiering completo dei file

Anche se il tiering FabricPool opera a livello di blocco, in alcuni casi può essere utilizzato per fornire un tiering a livello di file.

Molti set di dati delle applicazioni sono organizzati per data e tali dati hanno generalmente sempre meno probabilità di accedere man mano che invecchiano. Ad esempio, una banca potrebbe avere un archivio di file PDF che contengono cinque anni di dichiarazioni dei clienti, ma solo gli ultimi mesi sono attivi. FabricPool può essere utilizzato per spostare i file di dati meno recenti nel Tier di capacità. Un periodo di raffreddamento di 14 giorni garantirebbe che i 14 giorni più recenti di file PDF rimangano sul livello di prestazioni. Inoltre, i file letti almeno ogni 14 giorni resterebbero hot e quindi nel Tier di performance.

Policy

Per implementare un approccio di tiering basato su file, è necessario disporre di file scritti e non modificati successivamente. Il `tiering-minimum-cooling-days` i criteri devono essere impostati su un livello sufficientemente alto da mantenere i file di cui potresti aver bisogno nel tier di performance. Ad esempio, un set di dati per il quale sono necessari gli ultimi 60 giorni di dati con performance ottimali garantisce la definizione di `tiering-minimum-cooling-days` periodo a 60. Risultati simili possono essere ottenuti anche in base ai modelli di accesso ai file. Ad esempio, se sono necessari gli ultimi 90 giorni di dati e l'applicazione sta accedendo a quell'arco di dati di 90 giorni, i dati resteranno sul Tier di performance. Impostando `tiering-minimum-cooling-days` a 2, si ottiene un tiering prompt dopo che i dati sono meno attivi.

Il `auto` la policy è necessaria per gestire il tiering di questi blocchi perché solo l' `auto` il criterio influisce sui blocchi che si trovano nel file system attivo.

Qualsiasi tipo di accesso ai dati ripristina i dati della mappa termica. La scansione virus, l'indicizzazione e persino le attività di backup in grado di leggere i file di origine impediscono il tiering perché è necessario `tiering-minimum-cooling-days` la soglia non viene mai raggiunta.

Tiering parziale dei file

Poiché FabricPool opera a livello di blocchi, i file soggetti a modifiche possono essere parzialmente suddivisi in Tier nello storage a oggetti e rimanere parzialmente anche nel Tier di performance.

Ciò è comune con i database. Anche i database che contengono blocchi inattivi sono candidati per il tiering FabricPool. Ad esempio, un database di gestione della catena logistica potrebbe contenere informazioni cronologiche che devono essere disponibili se necessario ma non accessibili durante le normali operazioni. La funzione FabricPool può essere utilizzata per spostare selettivamente i blocchi inattivi.

Ad esempio, i file di dati in esecuzione su un volume FabricPool con un `tiering-minimum-cooling-days` il periodo di 90 giorni conserva i blocchi a cui si accede nei 90 giorni precedenti nel tier di performance. Tuttavia, qualsiasi elemento a cui non si accede per 90 giorni viene ricollocato nel Tier di capacità. In altri casi, la normale attività applicativa preserva i blocchi corretti sul livello corretto. Ad esempio, se un database viene normalmente utilizzato per elaborare regolarmente i 60 giorni precedenti di dati, è molto più basso `tiering-minimum-cooling-days` il periodo può essere impostato perché l'attività naturale dell'applicazione garantisce che i blocchi non vengano spostati prematuramente.

Il `auto` i criteri devono essere utilizzati con attenzione per i database. Numerosi database prevedono attività periodiche come la fine del quarter o la reindicizzazione delle operazioni. Se il periodo di queste operazioni è superiore a. `tiering-minimum-cooling-days` possono verificarsi problemi di prestazioni. Ad esempio, se l'elaborazione a fine quarter richiede 1TB TB di dati che non vengono intatti, è possibile che tali dati siano presenti nel Tier di capacità. Le letture dal Tier di capacità sono spesso estremamente veloci e potrebbero non causare problemi di performance, ma i risultati esatti dipendono dalla configurazione dell'archivio di oggetti.

Policy

Il `tiering-minimum-cooling-days` il criterio deve essere impostato su un livello sufficientemente alto da conservare i file che potrebbero essere necessari nel livello di prestazioni. Ad esempio, un database in cui potrebbero essere necessari gli ultimi 60 giorni di dati con prestazioni ottimali giustificherebbe l'impostazione di `tiering-minimum-cooling-days` periodo a 60 giorni. Risultati simili possono essere ottenuti anche in base ai modelli di accesso dei file. Ad esempio, se sono necessari gli ultimi 90 giorni di dati e l'applicazione sta accedendo a quell'arco di dati di 90 giorni, i dati resteranno sul Tier di performance. Impostazione di `tiering-minimum-cooling-days` un periodo di 2 giorni eseguirebbe il tiering dei dati non appena i dati diventano meno attivi.

Il `auto` la policy è necessaria per gestire il tiering di questi blocchi perché solo l' `auto` il criterio influisce sui blocchi che si trovano nel file system attivo.

Qualsiasi tipo di accesso ai dati ripristina i dati della mappa termica. Pertanto, le scansioni delle tabelle complete dei database e persino le attività di backup in grado di leggere i file di origine impediscono il tiering perché necessario `tiering-minimum-cooling-days` la soglia non viene mai raggiunta.

Tiering dei log di archivio

Forse l'utilizzo più importante per FabricPool è il miglioramento dell'efficienza dei dati cold noti, come i log delle transazioni dei database.

La maggior parte dei database relazionali opera in modalità di archiviazione dei log delle transazioni per fornire un ripristino point-in-time. Le modifiche apportate ai database vengono salvate registrando le modifiche nei registri delle transazioni e il registro delle transazioni viene conservato senza essere sovrascritto. Il risultato può essere la necessità di conservare un enorme volume di registri delle transazioni archiviati. Esempi simili esistono con molti altri flussi di lavoro delle applicazioni che generano dati che devono essere conservati, ma con molte probabilità di accesso.

FabricPool risolve questi problemi offrendo una singola soluzione con tiering integrato. I file vengono memorizzati e rimangono accessibili nella loro posizione abituale, ma non occupano praticamente spazio nell'array primario.

Policy

Utilizzare un `tiering-minimum-cooling-days` la policy di pochi giorni comporta la conservazione dei blocchi nei file creati di recente (che sono i file più probabilmente necessari a breve termine) nel tier di performance. I blocchi di dati dei file meno recenti vengono quindi spostati nel Tier di capacità.

Il `auto` applica il tiering prompt quando viene raggiunta la soglia di raffreddamento, indipendentemente dal fatto che i log siano stati eliminati o continuino a esistere nel file system primario. Inoltre, l'archiviazione di tutti i log potenzialmente necessari in un'unica posizione nel file system attivo semplifica la gestione. Non c'è motivo di cercare tra gli snapshot per individuare un file che deve essere ripristinato.

Alcune applicazioni, come Microsoft SQL Server, troncano i file di log delle transazioni durante le operazioni di backup in modo che i log non si trovino più nel file system attivo. È possibile risparmiare capacità utilizzando `snapshot-only` tiering delle policy, ma `auto` il criterio non è utile per i dati di log perché raramente dovrebbero essere raffreddati i dati di log nel file system attivo.

Tiering delle Snapshot

La release iniziale di FabricPool era rivolta a un caso di utilizzo di backup. L'unico tipo di blocchi che è possibile eseguire il tiering era costituito da blocchi che non erano più associati a dati nel file system attivo. Pertanto, solo i blocchi di dati Snapshot possono essere spostati nel Tier di capacità. Questa rimane una delle opzioni di tiering più sicure quando occorre, in modo da garantire che le performance non subiscano alcun impatto.

Criteri - istantanee locali

Esistono due opzioni per il tiering di blocchi di snapshot inattivi nel Tier di capacità. Innanzitutto, la `snapshot-only` la politica riguarda solo i blocchi di snapshot. Anche se il `auto` il criterio include `snapshot-only` ed esegue il tiering dei blocchi dal file system attivo. Ciò potrebbe non essere desiderabile.

Il `tiering-minimum-cooling-days` valore deve essere impostato su un periodo di tempo in cui i dati che potrebbero essere necessari durante un ripristino sono disponibili sul livello di prestazioni. Ad esempio, la maggior parte degli scenari di ripristino di un database di produzione critico include un punto di ripristino in un determinato momento dei giorni precedenti. Impostazione a. `tiering-minimum-cooling-days` il valore 3 garantisce che qualsiasi ripristino del file porti a un file che offre immediatamente le massime prestazioni. Tutti i blocchi dei file attivi sono ancora presenti sullo storage veloce senza dover ripristinarli dal livello di capacità.

Criteri - istantanee replicate

Di norma, uno snapshot replicato con SnapMirror o SnapVault utilizzato solo per il ripristino deve utilizzare FabricPool `all` policy. Con questa policy, i metadati vengono replicati, ma tutti i blocchi di dati vengono inviati immediatamente al Tier di capacità, ottenendo il massimo delle performance. La maggior parte dei processi di recovery implica un i/o sequenziale, che è intrinsecamente efficiente. È necessario valutare il tempo di ripristino dalla destinazione dell'archivio oggetti, ma in un'architettura ben progettata questo processo di ripristino non deve essere significativamente più lento del ripristino da dati locali.

Se per il cloning è prevista anche l'utilizzo dei dati replicati, l' `auto` la politica è più appropriata, con un `tiering-minimum-cooling-days` valore che comprende i dati che si prevede vengano utilizzati regolarmente in un ambiente di clonazione. Ad esempio, il working set attivo di un database potrebbe includere dati letti o scritti nei tre giorni precedenti, ma potrebbe includere anche altri 6 mesi di dati storici. In tal caso, il `auto` La policy nella destinazione di SnapMirror rende disponibile il working set nel Tier di performance.

Tiering del backup

I backup delle applicazioni tradizionali includono prodotti come Oracle Recovery Manager, che creano backup basati su file al di fuori della posizione del database originale.

```
 `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.
Il `auto` il criterio è il criterio più appropriato per i dati di backup.
In questo modo si garantisce un tiering rapido quando la soglia di
raffreddamento è stata raggiunta, indipendentemente dal fatto che i file
siano stati eliminati o continuino a esistere nel file system primario.
Inoltre, l'archiviazione di tutti i file potenzialmente necessari in
un'unica posizione nel file system attivo semplifica la gestione. Non c'è
motivo di cercare tra gli snapshot per individuare un file che deve essere
ripristinato.
```

Il `snapshot-only` i criteri potrebbero funzionare, ma si applicano solo ai blocchi che non si trovano più nel file system attivo. Pertanto, i file presenti in una condivisione NFS o SMB devono essere eliminati prima del tiering dei dati.

Questa policy risulterebbe ancora meno efficiente con la configurazione LUN, poiché l'eliminazione di un file da una LUN rimuove solo i riferimenti dei file dai metadati del file system. I blocchi effettivi sui LUN restano in posizione fino a quando non vengono sovrascritti. Questa situazione può creare un lungo ritardo tra il tempo di eliminazione di un file e il tempo in cui i blocchi vengono sovrascritti e candidati per il tiering. Lo spostamento dell' comporta alcuni vantaggi `snapshot-only` Dei blocchi nel Tier di capacità, ma, nel complesso, la gestione FabricPool dei dati di backup funziona meglio con l' `auto` policy.

Questo approccio aiuta gli utenti a gestire lo spazio richiesto per i backup in modo più efficiente, ma FabricPool non è una tecnologia di backup. Il tiering dei file di backup nell'archivio di oggetti semplifica la gestione perché i file sono ancora visibili nel sistema di storage originale, ma i blocchi di dati nella destinazione dell'archivio di oggetti dipendono dal sistema di storage originale. Se il volume di origine viene perso, i dati dell'archivio di oggetti non sono più utilizzabili.

Interruzioni di accesso agli archivi di oggetti

Il tiering di un set di dati con FabricPool determina una dipendenza tra lo storage array primario e il Tier dell'archivio di oggetti. Sono disponibili molte opzioni di storage a oggetti che offrono livelli di disponibilità variabili. È importante comprendere l'impatto di una possibile perdita di connettività tra lo storage array primario e il Tier dello storage a oggetti.

Se un i/o emesso a ONTAP richiede dati dal Tier di capacità e ONTAP non riesce a raggiungere il Tier di capacità per recuperare i blocchi, l'i/o finisce il time-out. L'effetto di questo timeout dipende dal protocollo utilizzato. In un ambiente NFS, ONTAP risponde con una risposta EJUKEBOX o EDELAY, a seconda del protocollo. Alcuni sistemi operativi meno recenti potrebbero interpretare questo come un errore, ma i sistemi operativi attuali e i livelli di patch correnti del client Oracle Direct NFS considerano questo come un errore recuperabile e continuano ad attendere il completamento dell'i/O.

Un timeout più breve si applica agli ambienti SAN. Se un blocco nell'ambiente dell'archivio oggetti è necessario e rimane irraggiungibile per due minuti, viene restituito un errore di lettura all'host. Il volume e i LUN di ONTAP rimangono online, ma il sistema operativo host potrebbe segnalare il file system come in uno stato di errore.

Problemi di connettività dello storage a oggetti `snapshot-only` i criteri sono meno preoccupanti, perché vengono suddivisi in livelli solo i dati di backup. I problemi di comunicazione rallenterebbero il recupero dei dati, ma non influenzerebbero altrimenti l'utilizzo attivo dei dati. Il `auto` e. `all` Le policy consentono il tiering dei dati cold dal LUN attivo, il che significa che un errore durante il recupero dei dati dell'archivio oggetti può influire sulla disponibilità del database. Un'implementazione SAN con queste policy deve essere utilizzata solo con storage a oggetti di classe Enterprise e connessioni di rete progettate per l'alta disponibilità. NetApp StorageGRID è l'opzione superiore.

Data Protection Oracle

Data Protection con ONTAP

NetApp sa che i dati più mission-critical sono presenti nei database.

Un'azienda non può operare senza accesso ai propri dati, e a volte i dati definiscono l'azienda. Questi dati devono essere protetti; tuttavia, la protezione dei dati non è solo garanzia di un backup utilizzabile, ma consiste nell'eseguire i backup in modo rapido e affidabile, oltre a memorizzarli in modo sicuro.

L'altro lato della protezione dei dati è la recovery. Quando i dati sono inaccessibili, l'azienda ne è interessata e potrebbe non funzionare fino a quando i dati non vengono ripristinati. Questo processo deve essere rapido e affidabile. Infine, la maggior parte dei database deve essere protetta dai disastri, il che significa mantenere una replica del database. La replica deve essere sufficientemente aggiornata. Rendere la replica un database completamente operativo deve anche essere semplice e veloce.

Questa documentazione sostituisce il report tecnico precedentemente pubblicato *TR-4591: Data Protection di Oracle: Backup, recovery e replica.*

Pianificazione

La corretta architettura di protezione dei dati aziendali dipende dai requisiti di business correlati a conservazione dei dati, ripristinabilità e tolleranza per le interruzioni durante i vari eventi.

Ad esempio, consideriamo il numero di applicazioni, database e set di dati importanti inclusi nell'ambito della fornitura. La costruzione di una strategia di backup per un singolo set di dati che garantisca la conformità con gli SLA tipici è piuttosto semplice, perché non ci sono molti oggetti da gestire. Con l'aumento del numero di set di dati, il monitoraggio diventa più complicato e gli amministratori potrebbero essere costretti a spendere una crescente quantità di tempo nella risoluzione degli errori di backup. Quando un ambiente raggiunge il cloud e scala un service provider, è necessario un approccio completamente diverso.

Anche le dimensioni del set di dati influiscono sulla strategia. Ad esempio, esistono molte opzioni per il backup e il ripristino con un database da 100GB TB perché il set di dati è così piccolo. La semplice copia dei dati dai supporti di backup con gli strumenti tradizionali in genere offre un RTO sufficiente per il recovery. Un database 100TB ha normalmente bisogno di una strategia completamente diversa, a meno che l'RTO non consenta un'interruzione di più giorni, nel qual caso una tradizionale procedura di backup e ripristino basata sulla copia potrebbe essere accettabile.

Infine, vi sono alcuni fattori che esulano dal processo di backup e ripristino stesso. Ad esempio, esistono database che supportano attività di produzione critiche e che rendono il ripristino una rara eventualità che viene eseguita solo da DBA esperti? In alternativa, i database fanno parte di un grande ambiente di sviluppo in cui il ripristino è un evento frequente e gestito da un team IT generico?

RTO, RPO e pianificazione SLA

ONTAP ti consente di personalizzare facilmente una strategia di protezione dei dati dei database di Oracle in base ai tuoi requisiti di business.

Questi requisiti includono fattori quali la velocità del recovery, la perdita massima consentita di dati e le esigenze di conservazione del backup. Il piano di protezione dei dati deve anche tenere in considerazione i vari requisiti normativi per la conservazione e il ripristino dei dati. Infine, è necessario considerare diversi scenari di ripristino dei dati, che vanno dal recupero tipico e prevedibile derivante da errori di utenti o applicazioni fino a scenari di ripristino di emergenza che includono la perdita completa di un sito.

Piccole modifiche alle policy di protezione e ripristino dei dati possono avere un effetto significativo sull'architettura generale dello storage, del backup e del ripristino. È fondamentale definire e documentare gli standard prima di iniziare il lavoro di progettazione, per evitare di complicare un'architettura di protezione dati. Le funzioni o i livelli di protezione non necessari comportano costi e costi di gestione inutili, mentre un requisito inizialmente trascurato può condurre un progetto nella direzione sbagliata o richiedere modifiche di progettazione dell'ultimo minuto.

Recovery time objective

L'obiettivo RTO (Recovery Time Objective) definisce il tempo massimo consentito per il ripristino di un servizio. Ad esempio, un database di risorse umane potrebbe avere un RTO di 24 ore perché, sebbene sarebbe molto scomodo perdere l'accesso a questi dati durante la giornata lavorativa, l'azienda può comunque operare. Al contrario, un database che supporta la contabilità generale di una banca avrebbe un RTO misurato in minuti o anche secondi. Un RTO di zero non è possibile, perché deve esserci un modo per distinguere tra un'effettiva interruzione del servizio e un evento di routine, come un pacchetto di rete perso. Tuttavia, un RTO prossimo

allo zero è un requisito tipico.

Obiettivo RPO

Il recovery point objective (RPO) definisce la massima perdita di dati tollerabile. In molti casi, l'RPO è determinato unicamente dalla frequenza delle snapshot o degli aggiornamenti di snapmirror.

In alcuni casi, l'RPO può essere reso più aggressivo proteggendo determinati dati con maggiore frequenza. In un contesto di database, l'RPO è in genere una questione di quanti dati di registro possono essere persi in una situazione specifica. In uno scenario di ripristino tipico in cui un database viene danneggiato a causa di un bug del prodotto o di un errore dell'utente, l'RPO deve essere pari a zero, il che significa che non ci devono essere perdite di dati. La procedura di ripristino prevede il ripristino di una copia precedente dei file di database e la riproduzione dei file di registro per portare lo stato del database al momento desiderato. I file di registro necessari per questa operazione dovrebbero essere già presenti nella posizione originale.

In scenari insoliti, i dati del registro potrebbero andare persi. Ad esempio, un accidentale o dannoso `rm -rf *` di file di database potrebbe causare l'eliminazione di tutti i dati. L'unica opzione sarebbe il ripristino dal backup, inclusi i file di registro, e alcuni dati andrebbero inevitabilmente persi. L'unica opzione per migliorare gli RPO in un ambiente di backup tradizionale sarebbe l'esecuzione di backup ripetuti dei dati di log. Questo comporta dei limiti, tuttavia, a causa dello spostamento costante dei dati e della difficoltà di mantenere un sistema di backup come servizio in esecuzione costante. Uno dei benefici dei sistemi storage avanzati è la capacità di proteggere i dati da danni accidentali o dannosi ai file e garantire quindi un RPO migliore senza spostamento dei dati.

Disaster recovery

Il ripristino di emergenza include l'architettura IT, i criteri e le procedure necessarie per il ripristino di un servizio in caso di emergenza fisica. Tra questi, inondazioni, incendi o persone che agiscono con intento doloso o negligente.

Il disaster recovery non è solo una serie di procedure di ripristino. Si tratta del processo completo di identificazione dei vari rischi, definizione dei requisiti di ripristino dei dati e continuità del servizio e realizzazione della giusta architettura con le relative procedure.

Durante la definizione dei requisiti di protezione dei dati, è fondamentale differenziare tra i requisiti tipici di RPO e RTO e quelli di RPO e RTO necessari per il disaster recovery. Alcuni ambienti applicativi richiedono un RPO pari a zero e un RTO prossimo allo zero per situazioni di perdita di dati che vanno da errori utente relativamente normali a incendi che distruggono un data center. Tuttavia, vi sono conseguenze amministrative e di costo per questi elevati livelli di protezione.

In generale, i requisiti di ripristino dei dati non di emergenza devono essere rigorosi per due motivi. Innanzitutto, i bug applicativi e gli errori degli utenti che danneggiano i dati sono prevedibili al punto che sono quasi inevitabili. In secondo luogo, non è difficile progettare una strategia di backup in grado di offrire un RPO pari a zero e un RTO basso finché il sistema storage non viene distrutto. Non c'è motivo di non affrontare un rischio significativo che sia facilmente risolvibile, motivo per cui gli obiettivi di RPO e RTO per il ripristino locale dovrebbero essere aggressivi.

I requisiti di RTO e RPO per il disaster recovery variano in modo più ampio in base alla probabilità di un disastro e alle conseguenze della perdita di dati associata o dell'interruzione di un business. I requisiti di RPO e RTO devono essere basati sulle effettive esigenze di business e non su principi generali. Devono tenere conto di più scenari di emergenza logici e fisici.

Disastri logici

I disastri logici includono la corruzione dei dati causata da utenti, bug delle applicazioni o del sistema operativo e malfunzionamenti del software. I disastri logici possono includere anche attacchi dannosi da parte di terzi

con virus o worm o sfruttando le vulnerabilità delle applicazioni. In questi casi, l'infrastruttura fisica rimane intatta, ma i dati sottostanti non sono più validi.

Un tipo sempre più comune di disastro logico è noto come ransomware, in cui un vettore di attacco viene utilizzato per crittografare i dati. La crittografia non danneggia i dati, ma li rende non disponibili fino a quando non viene effettuato il pagamento a terzi. Un numero sempre crescente di aziende è specificatamente preso di mira con gli hack ransomware. A causa di questa minaccia, NetApp offre snapshot a prova di manomissione, in cui nemmeno l'amministratore dello storage può modificare i dati protetti prima della data di scadenza configurata.

Disastri fisici

I disastri fisici includono l'errore di componenti di un'infrastruttura che superano le sue capacità di ridondanza e causano una perdita di dati o un'estesa perdita di servizio. Ad esempio, la protezione RAID fornisce la ridondanza dell'unità disco e l'utilizzo di HBA fornisce la ridondanza di porte FC e cavi FC. I guasti hardware di tali componenti sono prevedibili e non influiscono sulla disponibilità.

In un ambiente aziendale, è generalmente possibile proteggere l'infrastruttura di un intero sito con componenti ridondanti fino al punto in cui l'unico scenario di emergenza fisica prevedibile è la perdita completa del sito. Quindi, il piano di disaster recovery dipende dalla replica sito-sito.

Protezione dei dati sincrona e asincrona

In un mondo ideale, tutti i dati verrebbero replicati in modo sincrono tra siti dispersi geograficamente. Tale replicazione non è sempre fattibile o addirittura possibile per diversi motivi:

- La replica sincrona aumenta inevitabilmente la latenza di scrittura, perché tutte le modifiche devono essere replicate in entrambe le posizioni prima che l'applicazione/database possa procedere con l'elaborazione. L'effetto sulle prestazioni risultante è talvolta inaccettabile, escludendo l'uso del mirroring sincrono.
- La maggiore adozione di storage SSD al 100% implica maggiore probabilità di ottenere una latenza di scrittura aggiuntiva, poiché le aspettative di performance includono centinaia di migliaia di IOPS e latenza sotto al millisecondo. Ottenere tutti i benefici dell'utilizzo di SSD al 100% può richiedere la revisione della strategia di disaster recovery.
- I set di dati continuano a crescere in termini di byte, creando difficoltà per garantire una larghezza di banda sufficiente a sostenere la replica sincrona.
- I set di dati crescono anche in termini di complessità, creando problemi con la gestione della replica sincrona su larga scala.
- Le strategie basate sul cloud spesso implicano maggiori distanze di replica e latenza, precludendo ulteriormente l'utilizzo di mirroring sincrono.

NetApp offre soluzioni che includono replica sincrona per le più esigenti richieste di recovery di dati e soluzioni asincrone che consentono performance e flessibilità migliori. Inoltre, la tecnologia NetApp si integra perfettamente con molte soluzioni di replica di terze parti, come Oracle DataGuard

Tempo di conservazione

L'ultimo aspetto di una strategia di protezione dei dati è il tempo di conservazione dei dati, che può variare drasticamente.

- Un requisito tipico è rappresentato da 14 giorni di backup notturni sul sito primario e 90 giorni di backup memorizzati su un sito secondario.
- Molti clienti creano archivi trimestrali autonomi archiviati su supporti diversi.

- Un database costantemente aggiornato potrebbe non richiedere i dati storici e i backup devono essere conservati solo per alcuni giorni.
- I requisiti normativi potrebbero richiedere la possibilità di recupero fino al punto in cui avviene una transazione arbitraria nell'arco di 365 giorni.

Disponibilità del database

ONTAP è progettato per garantire la massima disponibilità dei database Oracle. Una descrizione completa delle funzioni di alta disponibilità di ONTAP esula dall'ambito di questo documento. Tuttavia, come per la protezione dei dati, una conoscenza di base di questa funzionalità è importante quando si progetta un'infrastruttura di database.

Coppie HA

L'unità di base dell'alta disponibilità è la coppia ha. Ciascuna coppia contiene collegamenti ridondanti per supportare la replica dei dati nella NVRAM. NVRAM non è una cache di scrittura. La RAM all'interno del controller funge da cache di scrittura. Lo scopo della NVRAM è quello di memorizzare temporaneamente i dati come salvaguardia da errori di sistema imprevisti. A questo proposito, è simile a un log di ripristino del database.

Sia la NVRAM che il redo log del database consentono di memorizzare i dati rapidamente, consentendo il commit delle modifiche ai dati il più rapidamente possibile. L'aggiornamento ai dati persistenti sulle unità (o file di dati) viene eseguito solo in un secondo momento durante un processo chiamato checkpoint sulle piattaforme ONTAP e sulla maggior parte dei database. Durante le normali operazioni, non vengono letti i dati della NVRAM né i log di ripristino del database.

Se un controller si guasta bruscamente, è probabile che vi siano modifiche in sospeso memorizzate nella NVRAM che non sono ancora state scritte sulle unità. Il partner controller rileva il guasto, assume il controllo dei dischi e applica le modifiche richieste che sono state memorizzate nella NVRAM.

Takeover e giveback

Il takeover e il giveback fanno riferimento al processo di trasferimento della responsabilità delle risorse di storage fra i nodi di una coppia ha. L'acquisizione e il giveback presentano due aspetti:

- Gestione della connettività di rete che consente l'accesso alle unità
- Gestione delle unità stesse

Le interfacce di rete che supportano il traffico CIFS e NFS sono configurate sia con una posizione home che di failover. Un takeover include lo spostamento delle interfacce di rete nella loro abitazione temporanea su un'interfaccia fisica situata sulla stessa subnet della posizione originale. Un giveback prevede lo spostamento delle interfacce di rete nelle posizioni originali. Il comportamento esatto può essere regolato come richiesto.

Le interfacce di rete che supportano i protocolli a blocchi SAN, come iSCSI e FC, non vengono ricollocate durante il takeover e lo giveback. È invece necessario eseguire il provisioning delle LUN attraverso percorsi che includano una coppia ha completa che si traduce in un percorso primario e un percorso secondario.

È possibile configurare anche percorsi aggiuntivi per controller aggiuntivi in modo da supportare la riallocazione dei dati tra i nodi di un cluster più grande, non facente parte del processo di ha.

Il secondo aspetto del takeover e dello sconto è il trasferimento della proprietà del disco. Il processo esatto dipende da diversi fattori, tra cui il motivo del takeover/giveback e le opzioni della riga di comando emesse.

L'obiettivo è quello di eseguire l'operazione nel modo più efficiente possibile. Anche se il processo complessivo potrebbe richiedere diversi minuti, il momento effettivo in cui la proprietà dell'unità viene trasferita da nodo a nodo può generalmente essere misurato in secondi.

Tempo di takeover

L'i/o dell'host subisce una breve pausa in i/o durante le operazioni di takeover e giveback, senza tuttavia alcuna interruzione dell'applicazione in un ambiente configurato correttamente. L'effettivo processo di transizione in cui l'i/o subisce un ritardo viene generalmente misurato in secondi, ma l'host potrebbe richiedere tempo aggiuntivo per riconoscere la modifica nei percorsi di dati e inviare di nuovo le operazioni i/O.

La natura dell'interruzione dipende dal protocollo:

- Un'interfaccia di rete che supporta il traffico NFS e CIFS emette una richiesta ARP (Address Resolution Protocol) alla rete dopo la transizione a una nuova posizione fisica. Ciò fa sì che gli switch di rete aggiornino le tabelle degli indirizzi MAC (Media Access Control) e riprendano l'elaborazione i/O. Le interruzioni nel caso di takeover e giveback pianificati vengono di solito misurate in secondi e in molti casi non sono rilevabili. Alcune reti potrebbero essere più lente a riconoscere completamente la modifica del percorso di rete e alcuni sistemi operativi potrebbero mettere in coda molti i/o in un breve periodo di tempo che deve essere rieseguito. Ciò può estendere il tempo necessario per riprendere l'i/O.
- Un'interfaccia di rete che supporta i protocolli SAN non passa a una nuova posizione. Un sistema operativo host deve modificare il percorso o i percorsi in uso. La pausa in i/o osservata dall'host dipende da diversi fattori. Dal punto di vista del sistema storage, il periodo in cui non è possibile fornire i/o è di pochi secondi. Tuttavia, sistemi operativi host diversi potrebbero richiedere tempo aggiuntivo per consentire un timeout i/o prima di riprovare. I sistemi operativi più recenti sono in grado di riconoscere un cambiamento di percorso molto più rapidamente, ma i sistemi operativi più vecchi in genere richiedono fino a 30 secondi per riconoscere un cambiamento.

La seguente tabella illustra i tempi di takeover previsti durante i quali il sistema storage non può fornire i dati a un ambiente applicativo. Non dovrebbero esserci errori in alcun ambiente applicativo, il takeover dovrebbe invece apparire come una breve pausa nell'elaborazione io.

	NFS	AFF	ASA
Takeover pianificato	15 sec.	6-10 sec.	2-3 sec.
Takeover non pianificato	30 sec.	6-10 sec.	2-3 sec.

Checksum e integrità dei dati

ONTAP e i protocolli supportati includono svariate funzionalità che proteggono l'integrità del database Oracle, inclusi dati a riposo e dati trasmessi sulla rete.

La protezione dei dati logici all'interno di ONTAP è costituita da tre requisiti principali:

- I dati devono essere protetti dalla corruzione.
- I dati devono essere protetti da guasti al disco.
- Le modifiche ai dati devono essere protette dalla perdita.

Queste tre esigenze sono discusse nelle sezioni seguenti.

Corruzione della rete: Checksum

Il livello più basilare di protezione dei dati è il checksum, che è uno speciale codice di rilevamento degli errori memorizzato insieme ai dati. La corruzione dei dati durante la trasmissione di rete viene rilevata con l'utilizzo di un checksum e, in alcuni casi, di checksum multipli.

Ad esempio, un frame FC include una forma di checksum chiamata CRC (Cyclic Redundancy Check) per assicurarsi che il payload non sia corrotto durante il transito. Il trasmettitore invia sia i dati che il CRC dei dati. Il ricevitore di un frame FC ricalcola il CRC dei dati ricevuti per assicurarsi che corrisponda al CRC trasmesso. Se il CRC appena calcolato non corrisponde al CRC collegato al frame, i dati sono corrotti e il frame FC viene scartato o rifiutato. Un'operazione i/o iSCSI include checksum ai livelli TCP/IP ed Ethernet e, per una maggiore protezione, può anche includere la protezione CRC opzionale al livello SCSI. Qualsiasi corruzione di bit sul filo viene rilevata dal livello TCP o IP, che porta alla ritrasmissione del pacchetto. Come nel caso di FC, gli errori nel CRC SCSI determinano un'eliminazione o un rifiuto dell'operazione.

Corruzione dei dischi: Checksum

I checksum vengono utilizzati anche per verificare l'integrità dei dati memorizzati sui dischi. I blocchi di dati scritti sui dischi vengono memorizzati con una funzione di checksum che produce un numero imprevedibile e legato ai dati originali. Quando i dati vengono letti dall'unità, il checksum viene ricalcolato e confrontato con il checksum memorizzato. Se non corrisponde, i dati sono corrotti e devono essere recuperati dal livello RAID.

Corruzione dei dati: Scritture perse

Uno dei tipi più difficili di corruzione da rilevare è una scrittura persa o posizionata erroneamente. Quando una scrittura viene confermata, deve essere scritta sul supporto nella posizione corretta. La corruzione dei dati sul posto è relativamente semplice da rilevare utilizzando un semplice checksum memorizzato con i dati. Tuttavia, se la scrittura viene semplicemente persa, la versione precedente dei dati potrebbe ancora esistere e il checksum sarebbe corretto. Se la scrittura viene posizionata nella posizione fisica errata, il checksum associato sarebbe ancora una volta valido per i dati memorizzati, anche se la scrittura ha distrutto altri dati.

La soluzione a questa sfida è la seguente:

- Un'operazione di scrittura deve includere metadati che indicano la posizione in cui dovrebbe essere trovata la scrittura.
- Un'operazione di scrittura deve includere un tipo di identificatore di versione.

Quando ONTAP scrive un blocco, include i dati sulla posizione di appartenenza del blocco. Se una lettura successiva identifica un blocco, ma i metadati indicano che esso appartiene alla posizione 123 quando è stato trovato nella posizione 456, allora la scrittura è stata erroneamente posizionata.

Rilevare una scrittura totalmente persa è più difficile. La spiegazione è molto complicata, ma essenzialmente ONTAP memorizza i metadati in modo che un'operazione di scrittura determini aggiornamenti a due posizioni diverse sulle unità. Se una scrittura viene persa, una successiva lettura dei dati e dei metadati associati mostra due diverse identità di versione. Ciò indica che la scrittura non è stata completata dall'unità.

La corruzione in scrittura persa e posizionata erroneamente è estremamente rara, ma con il continuo aumento dei dischi e la diminuzione dei set di dati nella scala di exabyte, il rischio aumenta. Il rilevamento delle operazioni di scrittura perse deve essere incluso in qualsiasi sistema storage che supporti i carichi di lavoro del database.

Guasti del disco: RAID, RAID DP e RAID-TEC

Se un blocco di dati su un'unità viene rilevato come danneggiato o se l'intera unità si guasta e non è completamente disponibile, i dati devono essere ricostituiti. Questo viene fatto in ONTAP utilizzando unità di

parità. Lo striping dei dati viene eseguito su più unità dati, quindi vengono generati i dati di parità. I dati vengono memorizzati separatamente dai dati originali.

ONTAP utilizzava in origine RAID 4, che utilizza un singolo disco di parità per ciascun gruppo di unità dati. Il risultato è che un'unità del gruppo potrebbe guastarsi senza causare una perdita di dati. Se l'unità di parità non funziona correttamente, non sono stati danneggiati dati ed è stato possibile costruire una nuova unità di parità. Se si è verificato un errore in una singola unità dati, è possibile utilizzare le unità rimanenti con l'unità di parità per rigenerare i dati mancanti.

Quando le unità erano di piccole dimensioni, la possibilità statistica di due unità che si guastavano contemporaneamente era trascurabile. Con la progressiva crescita della capacità del disco aumentano anche il tempo necessario per ricostruire i dati in seguito a un guasto al disco. Ciò ha aumentato la finestra in cui un guasto di una seconda unità causerebbe la perdita di dati. Inoltre, il processo di ricostruzione crea numerosi i/o aggiuntivi sui dischi ancora in uso. Man mano che i dischi diventano obsoleti, aumenta anche il rischio di carico aggiuntivo che potrebbe causare un guasto al secondo disco. Infine, anche se il rischio di perdita di dati non aumentasse con il continuo utilizzo di RAID 4, le conseguenze della perdita di dati diventerebbero più gravi. Maggiore è la quantità di dati che andrebbero persi in caso di guasto a un gruppo RAID, più tempo occorrerebbe per ripristinare i dati, prolungando l'interruzione del business.

Questi problemi hanno portato NetApp a sviluppare la tecnologia NetApp RAID DP, una variante di RAID 6. Questa soluzione include due unità di parità, il che significa che due unità in un gruppo RAID possono guastarsi senza creare perdite di dati. Le dimensioni dei dischi hanno continuato a crescere, portando infine NetApp a sviluppare la tecnologia NetApp RAID-TEC, che introduce un disco a terza parità.

Alcune procedure consigliate per i database storici consigliano l'uso di RAID-10, noto anche come mirroring con striping. Ciò offre una protezione dei dati inferiore rispetto a quella dei sistemi RAID DP, in quanto vi sono più scenari di guasto a due dischi, mentre in RAID DP non ve ne sono nessuno.

Esistono inoltre alcune procedure consigliate per i database storici che indicano che le opzioni RAID-10 sono preferite a quelle RAID-4/5/6 a causa di problemi di prestazioni. Queste raccomandazioni a volte fanno riferimento a una penalizzazione RAID. Sebbene queste raccomandazioni siano generalmente corrette, non sono applicabili alle implementazioni di RAID all'interno di ONTAP. Il problema di prestazioni è relativo alla rigenerazione di parità. Con le implementazioni RAID tradizionali, l'elaborazione delle random write di routine eseguite da un database richiede letture multiple del disco per rigenerare i dati di parità e completare la scrittura. La penalità viene definita come gli IOPS in lettura aggiuntivi necessari per eseguire le operazioni di scrittura.

ONTAP non subisce alcuna penalizzazione RAID perché le scritture vengono organizzate in memoria dove la parità viene generata e quindi scritta su disco come singolo stripe RAID. Non sono richieste letture per completare l'operazione di scrittura.

In sintesi, rispetto al RAID 10, RAID DP e RAID-TEC offrono una capacità utilizzabile molto maggiore, una migliore protezione contro i guasti ai dischi e nessun compromesso in termini di performance.

Protezione da errori hardware: NVRAM

Qualsiasi storage array che gestisce un carico di lavoro del database deve eseguire le operazioni di scrittura il più rapidamente possibile. Inoltre, un'operazione di scrittura deve essere protetta dalla perdita da un evento imprevisto, come un'interruzione dell'alimentazione. Ciò significa che qualsiasi operazione di scrittura deve essere conservata in modo sicuro in almeno due posizioni.

I sistemi AFF e FAS si affidano alla NVRAM per soddisfare questi requisiti. Il processo di scrittura funziona come segue:

1. I dati di scrittura in entrata sono memorizzati nella RAM.

2. Le modifiche che devono essere apportate ai dati sul disco vengono registrate nella NVRAM sia sul nodo locale che sul nodo partner. NVRAM non è una cache di scrittura, ma un journal simile a un log di ripristino dei database. In condizioni normali, non viene letta. Viene utilizzata solo per il ripristino, ad esempio in seguito a un'interruzione dell'alimentazione durante l'elaborazione i/O.
3. La scrittura viene quindi riconosciuta all'host.

Il processo di scrittura in questa fase è completo dal punto di vista dell'applicazione e i dati sono protetti dalla perdita, perché vengono memorizzati in due posizioni diverse. Alla fine, le modifiche vengono scritte su disco, ma il processo risulta fuori banda dal punto di vista dell'applicazione perché si verifica dopo il riconoscimento della scrittura e quindi non influisce sulla latenza. Questo processo è ancora una volta simile alla registrazione del database. Una modifica al database viene registrata nei registri di ripristino il più rapidamente possibile e la modifica viene quindi riconosciuta come confermata. Gli aggiornamenti ai file di dati avvengono molto più tardi e non influenzano direttamente la velocità di elaborazione.

In caso di guasto a un controller, il partner controller assume la proprietà dei dischi richiesti e riproduce i dati registrati nella NVRAM per ripristinare le operazioni di i/o in corso quando si è verificato il guasto.

Protezione da errori hardware: NVFAIL

Come discusso in precedenza, una scrittura non viene riconosciuta fino a quando non è stata registrata nella NVRAM locale e nella NVRAM su almeno un altro controller. Questo approccio garantisce che un guasto dell'hardware o un'interruzione di corrente non comporti la perdita dell'i/o in-flight In caso di guasto della NVRAM locale o di guasto della connettività al partner di ha, i dati in-flight non verranno più mirrorati.

Se la NVRAM locale riporta un errore, il nodo si arresta. Questo arresto determina il failover su un controller partner ha. Nessun dato viene perso perché il controller che presenta il guasto non ha confermato l'operazione di scrittura.

ONTAP non consente un failover quando i dati non sono sincronizzati, a meno che il failover non sia forzato. La forzatura di una modifica delle condizioni in questo modo riconosce che i dati potrebbero essere lasciati indietro nel controllore originale e che la perdita di dati è accettabile.

I database sono particolarmente vulnerabili al danneggiamento se un failover viene forzato perché mantengono grandi cache interne di dati su disco. In caso di failover forzato, le modifiche precedentemente riconosciute vengono effettivamente eliminate. Il contenuto dell'array di storage torna indietro nel tempo e lo stato della cache del database non riflette più lo stato dei dati su disco.

Per proteggere i dati da questa situazione, ONTAP consente di configurare i volumi per una protezione speciale contro gli errori della NVRAM. Quando attivato, questo meccanismo di protezione determina l'ingresso di un volume nello stato chiamato NVFAIL. Questo stato causa errori di i/o che causano l'arresto di un'applicazione in modo che non utilizzino dati obsoleti. I dati non devono essere persi perché qualsiasi scrittura riconosciuta deve essere presente sull'array di storage.

Solitamente, gli amministratori dovranno arrestare completamente gli host prima di riportare manualmente LUN e volumi in linea. Sebbene queste fasi possano comportare un certo lavoro, questo approccio è il modo più sicuro per garantire l'integrità dei dati. Non tutti i dati richiedono questa protezione, motivo per cui il comportamento di NVFAIL può essere configurato in base al volume.

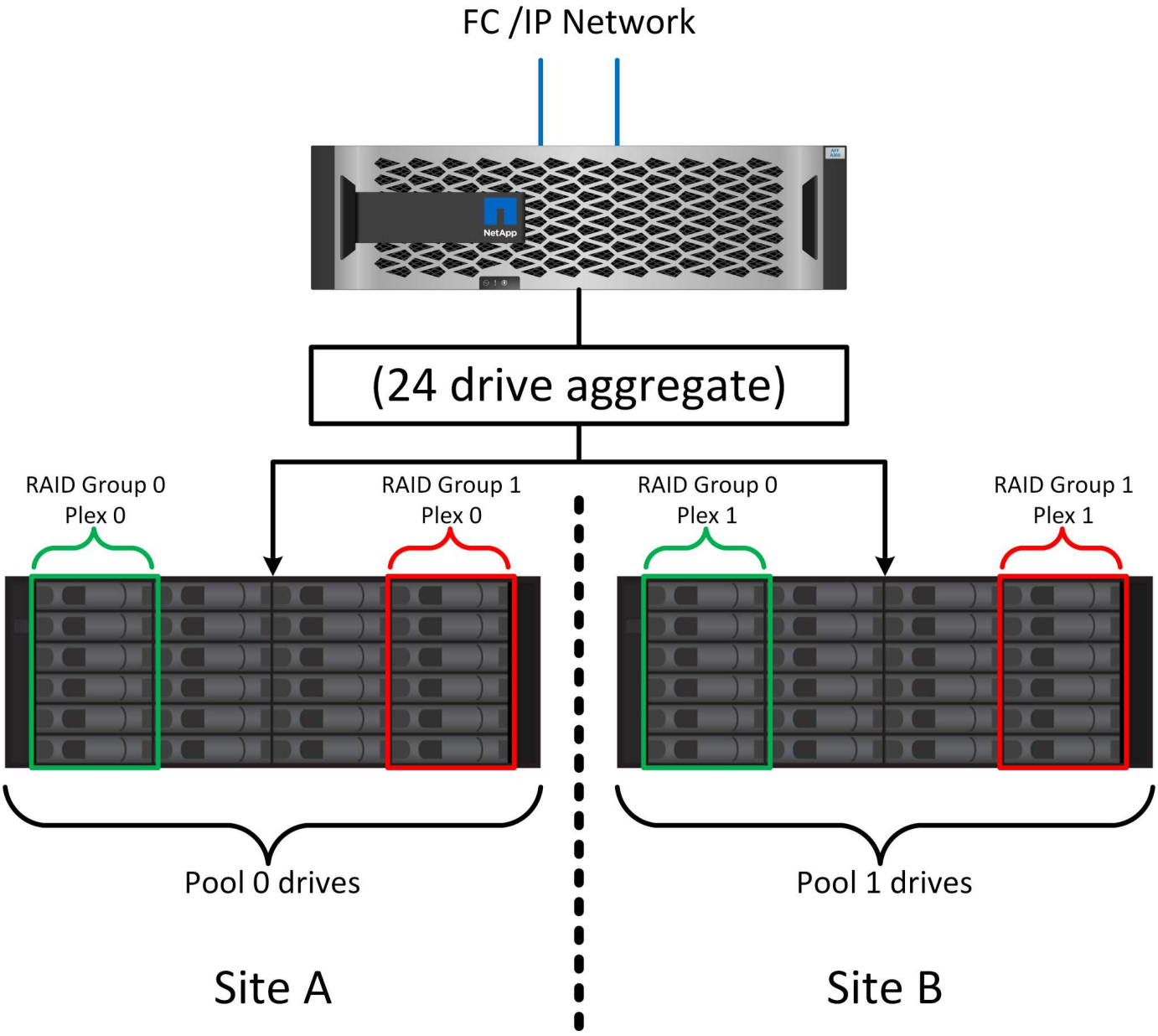
Protezione dai guasti di shelf e siti: SyncMirror e plessi

SyncMirror è una tecnologia di mirroring che migliora, ma non sostituisce, RAID DP o RAID-TEC. Esegue il mirroring del contenuto di due gruppi RAID indipendenti. La configurazione logica è la seguente:

- I dischi sono configurati in due pool in base alla posizione. Un pool è composto da tutti i dischi sul sito A,

mentre il secondo è composto da tutti i dischi sul sito B.

- Viene quindi creato un pool di storage comune, detto aggregato, in base a set di gruppi RAID con mirroring. Viene ottenuto lo stesso numero di unità per ciascun sito. Ad esempio, un aggregato SyncMirror da 20 dischi sarebbe composto da 10 dischi del sito A e 10 dischi del sito B.
- Ogni set di unità su un dato sito viene configurato automaticamente come uno o più gruppi RAID-DP o RAID-TEC completamente ridondanti, indipendentemente dall'utilizzo del mirroring. In questo modo si garantisce una protezione dei dati continua, anche dopo la perdita di un sito.



La figura precedente illustra una configurazione SyncMirror di esempio. È stato creato un aggregato di 24 dischi sul controller con 12 dischi da uno shelf allocato sul sito A e 12 dischi da uno shelf allocato sul sito B. I dischi sono stati raggruppati in due gruppi RAID con mirroring. Il gruppo RAID 0 include un plesso A 6 unità sul sito A con mirroring su un plesso A 6 unità sul sito B. Analogamente, il gruppo RAID 1 include un plesso A 6 unità sul sito A con mirroring su un plesso A 6 unità sul sito B.

Di norma, SyncMirror viene utilizzato per fornire il mirroring remoto con i sistemi MetroCluster, con una copia dei dati in ciascun sito. A volte, è stato utilizzato per fornire un livello di ridondanza extra in un unico sistema. In

particolare, fornisce ridondanza a livello di shelf. Uno shelf di dischi contiene già doppi controller e alimentatori e nel complesso è poco più di una lamiera, ma in alcuni casi è consigliabile garantire una protezione extra. Ad esempio, un cliente NetApp ha implementato SyncMirror per una piattaforma mobile di analytics in tempo reale utilizzata durante i test nel settore automobilistico. Il sistema è stato separato in due rack fisici forniti da alimentatori indipendenti da sistemi UPS indipendenti.

Checksum

L'argomento dei checksum è di particolare interesse per i DBA abituati all'utilizzo dei backup in streaming Oracle RMAN che migrano a backup basati su snapshot. Una caratteristica di RMAN è che esegue controlli di integrità durante le operazioni di backup. Sebbene questa funzionalità offra un certo valore, il suo vantaggio principale è quello di un database non utilizzato su uno storage array moderno. Quando si utilizzano dischi fisici per un database Oracle, è quasi certo che il danneggiamento si verifica anche in caso di invecchiamento dei dischi, un problema che viene risolto dai checksum basati su array negli storage array reali.

Con un vero storage array, l'integrità dei dati è protetta utilizzando checksum a livelli multipli. Se i dati sono corrotti in una rete basata su IP, il livello TCP (Transmission Control Protocol) rifiuta i dati a pacchetto e richiede la ritrasmissione. Il protocollo FC include i checksum, così come i dati SCSI incapsulati. Dopo essere stato inserito nell'array, ONTAP dispone della protezione RAID e checksum. Il danneggiamento può verificarsi, ma, come nella maggior parte degli array Enterprise, viene rilevato e corretto. In genere, si verifica un guasto di un intero disco, che richiede una ricostruzione RAID e l'integrità del database rimane inalterata. È ancora possibile che i singoli byte su un'unità siano danneggiati dalla radiazione cosmica o da celle flash difettose. In questo caso, il controllo della parità non viene eseguito correttamente, l'unità viene chiusa in errore e viene avviata la ricostruzione RAID. Ancora una volta, l'integrità dei dati non viene influenzata. L'ultima linea di difesa è l'uso di checksum. Se, ad esempio, un errore catastrofico del firmware su un'unità ha danneggiato i dati in un modo che in qualche modo non è stato rilevato da un controllo di parità RAID, il checksum non corrisponderebbe e ONTAP impedirebbe il trasferimento di un blocco danneggiato prima che il database Oracle potesse riceverlo.

L'architettura dei log di ripristino e file dati di Oracle è inoltre progettata per offrire il massimo livello di integrità dei dati possibile, anche in circostanze estreme. A livello massimo, i blocchi Oracle includono il checksum e controlli logici di base con quasi ogni i/O. Se Oracle non è in crash o non ha portato offline uno spazio di tabella, i dati saranno intatti. Il grado di controllo dell'integrità dei dati è regolabile e Oracle può anche essere configurato per confermare le operazioni di scrittura. Di conseguenza, è possibile ripristinare quasi tutti gli scenari di crash e di guasto e, nel caso estremamente raro di una situazione irreversibile, viene immediatamente rilevata la corruzione.

La maggior parte dei clienti NetApp che utilizzano database Oracle interrompe l'utilizzo di RMAN e di altri prodotti di backup dopo la migrazione a backup basati su snapshot. Esistono ancora opzioni in cui RMAN può essere utilizzato per eseguire un ripristino a livello di blocco con SnapCenter. Tuttavia, ogni giorno, RMAN, NetBackup e altri prodotti vengono utilizzati solo occasionalmente per creare copie di archivio mensili o trimestrali.

Alcuni clienti scelgono di eseguire `dbv` eseguire periodicamente controlli di integrità dei database esistenti. NetApp scoraggia questa pratica perché crea un carico i/o non necessario. Come illustrato in precedenza, se il database non presentava problemi, la possibilità di `dbv` Il rilevamento di un problema è prossimo allo zero e questa utility crea un carico i/o sequenziale molto elevato sulla rete e sul sistema di storage. A meno che non vi sia motivo di ritenere che esista una corruzione, come l'esposizione a un bug Oracle noto, non c'è motivo di eseguire `dbv`.

Elementi di base di backup e recovery

Backup basati su snapshot

La base della protezione dei dati dei database Oracle su ONTAP è la tecnologia Snapshot di NetApp.

I valori chiave sono i seguenti:

- **Semplicità.** Uno snapshot è una copia di sola lettura del contenuto di un contenitore di dati in un determinato momento.
- **Efficienza.** le istantanee non richiedono spazio al momento della creazione. Lo spazio viene occupato solo quando i dati vengono modificati.
- **Gestibilità.** Una strategia di backup basata sugli snapshot è facile da configurare e gestire perché gli snapshot sono parte nativa del sistema operativo di storage. Se il sistema di archiviazione è acceso, è pronto per creare dei backup.
- **Scalabilità.** è possibile conservare fino a 1024 backup di un singolo contenitore di file e LUN. Per set di dati complessi, più container di dati possono essere protetti da un singolo set coerente di snapshot.
- Le prestazioni non sono influenzate, indipendentemente dal fatto che un volume contenga 1024 snapshot o nessuno.

Sebbene molti vendor di soluzioni storage offrano la tecnologia Snapshot, la tecnologia Snapshot all'interno di ONTAP è unica e offre benefici significativi per gli ambienti applicativi aziendali e di database:

- Le copie snapshot fanno parte del layout file WAFL (Write-Anywhere file Layout) sottostante. Non si tratta di una tecnologia aggiuntiva o esterna. Questo semplifica la gestione, perché il sistema storage è il sistema di backup.
- Le copie Snapshot non influiscono sulle prestazioni, ad eccezione di alcuni casi edge, come ad esempio quando una quantità così elevata di dati viene memorizzata nelle snapshot che il sistema storage sottostante si riempie.
- Il termine "gruppo di coerenza" viene spesso utilizzato per fare riferimento a un raggruppamento di oggetti di storage che vengono gestiti come una raccolta coerente di dati. Uno snapshot di un particolare volume ONTAP costituisce il backup del gruppo di coerenza.

Le snapshot ONTAP offrono anche una scalabilità migliore rispetto alle tecnologie della concorrenza. I clienti possono memorizzare 5, 50 o 500 snapshot senza influire sulle performance. Il numero massimo di snapshot attualmente consentiti in un volume è 1024. Se è necessaria una conservazione aggiuntiva degli snapshot, sono disponibili opzioni per trasferire gli snapshot in cascata ad altri volumi.

Di conseguenza, la protezione di un set di dati ospitato su ONTAP è semplice e altamente scalabile. I backup non richiedono lo spostamento dei dati, pertanto una strategia di backup può essere personalizzata in base alle esigenze dell'azienda piuttosto che alle limitazioni delle velocità di trasferimento di rete, del numero elevato di unità a nastro o delle aree di staging del disco.

Uno snapshot è un backup?

Una domanda comunemente posta sull'utilizzo delle istantanee come strategia di protezione dei dati è il fatto che i dati "reali" e i dati snapshot si trovano sulle stesse unità. La perdita di tali unità causerebbe la perdita sia dei dati primari che del backup.

Si tratta di un problema valido. Le snapshot locali vengono utilizzate per le esigenze di backup e ripristino quotidiane, e in questo senso la snapshot è un backup. Quasi il 99% di tutti gli scenari di ripristino in ambienti NetApp si affida alle snapshot per soddisfare anche i requisiti RTO più aggressivi.

Gli snapshot locali, tuttavia, non dovrebbero mai rappresentare l'unica strategia di backup, motivo per cui NetApp offre tecnologie come SnapMirror e la replica SnapVault per replicare in modo rapido ed efficiente le snapshot su un set indipendente di dischi. In una soluzione adeguatamente progettata con istantanee e replica snapshot, l'utilizzo del nastro può essere ridotto a icona in un archivio trimestrale o eliminato del tutto.

Backup basati su snapshot

Le copie Snapshot di ONTAP sono disponibili diverse opzioni per la protezione dei dati, mentre le snapshot sono alla base di molte altre funzionalità di ONTAP, tra cui replica, disaster recovery e cloning. Una descrizione completa della tecnologia snapshot non rientra nell'ambito di questo documento, ma le sezioni seguenti forniscono una panoramica generale.

Esistono due approcci principali per creare uno snapshot di un dataset:

- Backup coerenti con il crash
- Backup coerenti con le applicazioni

Un backup coerente con i crash di un set di dati si riferisce all'acquisizione dell'intera struttura di set di dati in un singolo point-in-time. Se il set di dati è memorizzato in un singolo volume, il processo è semplice ed è possibile creare una Snapshot in qualsiasi momento. Se un set di dati si estende tra i volumi, è necessario creare uno snapshot del gruppo di coerenza (CG). Esistono diverse opzioni per la creazione di snapshot CG, tra cui il software NetApp SnapCenter, le funzionalità native del gruppo di coerenza ONTAP e gli script gestiti dagli utenti.

I backup coerenti con i crash vengono utilizzati principalmente quando è sufficiente un ripristino point-of-the-backup. Quando è richiesto un ripristino più granulare, sono in genere necessari backup coerenti con l'applicazione.

La parola "coerente" in "coerente con l'applicazione" è spesso un nome scorretto. Ad esempio, l'inserimento di un database Oracle in modalità di backup viene definito backup coerente con l'applicazione, ma i dati non vengono resi coerenti o disattivati in alcun modo. I dati continuano a cambiare durante il backup. Al contrario, la maggior parte dei backup di MySQL e Microsoft SQL Server disattivano i dati prima di eseguire il backup. VMware può o non può rendere certi file coerenti.

Gruppi di coerenza

Il termine "gruppo di coerenza" si riferisce alla capacità di un array di archiviazione di gestire più risorse di archiviazione come una singola immagine. Ad esempio, un database può essere composto da 10 LUN. L'array deve essere in grado di eseguire il backup, il ripristino e la replica delle 10 LUN in modo coerente. Il ripristino non è possibile se le immagini dei LUN non erano coerenti nel punto di backup. La replica di queste 10 LUN richiede che tutte le repliche siano perfettamente sincronizzate l'una con l'altra.

Il termine "gruppo di coerenza" non viene spesso utilizzato quando si parla di ONTAP perché la coerenza è sempre stata una funzione di base dell'architettura di volumi e aggregati all'interno di ONTAP. Molti altri storage array gestiscono LUN o file system come unità singole. Possono quindi essere configurati facoltativamente come "gruppo di coerenza" ai fini della protezione dei dati, ma questo è un passaggio aggiuntivo nella configurazione.

ONTAP è sempre stata in grado di acquisire immagini di dati coerenti locali e replicate. Anche se i vari volumi su un sistema ONTAP non vengono in genere formalmente descritti come un gruppo di coerenza, è proprio questo lo sono. Una snapshot di tale volume è un'immagine del gruppo di coerenza, il ripristino di tale snapshot è un ripristino di un gruppo di coerenza e sia SnapMirror che SnapVault offrono la replica di un gruppo di coerenza.

Snapshot di gruppo di coerenza

Le snapshot di gruppo di coerenza (cg-Snapshot) sono un'estensione della tecnologia Snapshot di base di ONTAP. Un'operazione Snapshot standard crea un'immagine coerente di tutti i dati all'interno di un singolo volume, ma a volte è necessario creare un set coerente di Snapshot su più volumi e persino su sistemi di storage multipli. Ne risulta una serie di snapshot che possono essere utilizzate allo stesso modo di uno snapshot di un solo volume. Possono essere utilizzati per il recovery locale dei dati, replicati a scopo di disaster recovery o clonati come una singola unità coerente.

Il più grande utilizzo noto di cg-snapshot è per un ambiente di database di circa 1PB GB su 12 controller. Le cg-Snapshot create su questo sistema sono state utilizzate per il backup, il ripristino e il cloning.

Nella maggior parte dei casi, quando un set di dati copre i volumi e l'ordine di scrittura deve essere preservato, il software di gestione scelto utilizza automaticamente uno snapshot cg. In questi casi non è necessario comprendere i dettagli tecnici delle istantanee cg. Tuttavia, in alcune situazioni, i complessi requisiti di protezione dei dati richiedono un controllo dettagliato sul processo di protezione e replica dei dati. I flussi di lavoro di automazione o l'uso di script personalizzati per richiamare le API cg-snapshot sono alcune delle opzioni disponibili. La comprensione dell'opzione migliore e del ruolo di cg-snapshot richiede una spiegazione più dettagliata della tecnologia.

La creazione di una serie di istantanee cg è un processo in due fasi:

1. Stabilire il recencing in scrittura su tutti i volumi di destinazione.
2. Creare Snapshot di tali volumi nello stato fenced (fenced).

La recinzione in scrittura viene stabilita in serie. Ciò significa che, mentre il processo di scherma viene configurato su più volumi, l'i/o in scrittura viene bloccato sul primo volume della sequenza mentre continua ad essere assegnato ai volumi che compaiono in seguito. Questo potrebbe inizialmente sembrare una violazione del requisito per il mantenimento dell'ordine di scrittura, ma ciò si applica solo all'i/o emesso in modo asincrono sull'host e non dipende da altre scritture.

Ad esempio, un database potrebbe eseguire numerosi aggiornamenti asincroni del file dati, consentendo al sistema operativo di riordinare l'i/o e completarli in base alla propria configurazione dell'utilità di pianificazione. L'ordine di questo tipo di i/o non può essere garantito perché l'applicazione e il sistema operativo hanno già rilasciato il requisito di mantenere l'ordine di scrittura.

Come esempio di contatore, la maggior parte delle attività di registrazione del database è sincrona. Il database non procede con ulteriori scritture di registro fino a quando l'i/o non viene riconosciuto e l'ordine di tali scritture deve essere conservato. Se un i/o di registro arriva su un volume fenced, non viene riconosciuto e le applicazioni vengono bloccate in ulteriori scritture. Analogamente, l'i/o di metadati del file system è di solito sincrono. Ad esempio, un'operazione di eliminazione file non deve essere persa. Se un sistema operativo con un file system xfs eliminava un file e l'i/o che aggiornava i metadati del file system xfs per rimuovere il riferimento a quel file apposto su un volume recintato, l'attività del file system si interrompeva. Ciò garantisce l'integrità del file system durante le operazioni cg-snapshot.

Dopo aver configurato la funzionalità write fencing nei volumi di destinazione, sono pronti per la creazione di snapshot. Non è necessario creare esattamente gli snapshot contemporaneamente, perché lo stato dei volumi è bloccato da un punto di vista di scrittura dipendente. Per evitare un difetto nell'applicazione che crea le istantanee cg, la recinzione iniziale include un timeout configurabile in cui ONTAP rilascia automaticamente la recinzione e riprende l'elaborazione di scrittura dopo un numero definito di secondi. Se tutte le istantanee vengono create prima dello scadere del periodo di timeout, il gruppo risultante di istantanee è un gruppo di coerenza valido.

Ordine di scrittura dipendente

Da un punto di vista tecnico, la chiave per un gruppo di coerenza è preservare l'ordine di scrittura e, nello specifico, l'ordine di scrittura dipendente. Ad esempio, un database in scrittura su 10 LUN scrive simultaneamente su tutte. Molte scritture vengono emesse in modo asincrono, il che significa che l'ordine in cui vengono completate non è importante e l'ordine effettivo in cui vengono completate varia in base al comportamento del sistema operativo e della rete.

Alcune operazioni di scrittura devono essere presenti sul disco prima che il database possa procedere con operazioni di scrittura aggiuntive. Queste operazioni critiche di scrittura sono chiamate scritture dipendenti. I/o di scrittura successivi dipendono dalla presenza di queste scritture sul disco. Qualsiasi snapshot, recovery o replica di queste 10 LUN deve garantire l'ordine di scrittura dipendente. Gli aggiornamenti del file system sono un altro esempio di scritture dipendenti dall'ordine di scrittura. L'ordine in cui vengono apportate le modifiche al file system deve essere mantenuto o l'intero file system potrebbe danneggiarsi.

Strategie

Esistono due approcci principali ai backup basati su snapshot:

- Backup coerenti con il crash
- Backup a caldo protetti dagli snapshot

Un backup coerente con i crash di un database si riferisce all'acquisizione dell'intera struttura del database, inclusi i file di dati, i log di ripristino e i file di controllo, in un singolo momento. Se il database è memorizzato in un singolo volume, il processo è semplice ed è possibile creare una Snapshot in qualsiasi momento. Se un database si estende su volumi, è necessario creare uno snapshot del gruppo di coerenza (CG). Esistono diverse opzioni per la creazione di snapshot CG, tra cui il software NetApp SnapCenter, le funzionalità native del gruppo di coerenza ONTAP e gli script gestiti dagli utenti.

I backup Snapshot coerenti con i crash vengono utilizzati principalmente quando è sufficiente un recovery point-of-the-backup. In alcune circostanze è possibile applicare i registri di archivio, ma quando è necessario un ripristino point-in-time più granulare, è preferibile un backup online.

La procedura di base per un backup online basato su snapshot è la seguente:

1. Inserire il database in `backup` modalità.
2. Creare una snapshot di tutti i volumi che ospitano file di dati.
3. Esci `backup` modalità.
4. Eseguire il comando `alter system archive log current` per forzare l'archiviazione del registro.
5. Creare snapshot di tutti i volumi che ospitano i log di archivio.

Questa procedura produce una serie di istantanee contenenti file di dati in modalità backup e i registri di archivio critici generati in modalità backup. Questi sono i due requisiti per il ripristino di un database. I file come i file di controllo dovrebbero essere protetti per comodità, ma l'unico requisito assoluto è la protezione per i file di dati e i registri di archivio.

Sebbene i diversi clienti possano avere strategie molto diverse, quasi tutte queste strategie si basano in ultima analisi sugli stessi principi delineati di seguito.

Recovery basato su Snapshot

Quando si progettano layout di volumi per database Oracle, la prima decisione è se utilizzare la tecnologia VBSR (Volume-Based NetApp SnapRestore).

La funzione SnapRestore basata su volume consente di ripristinare quasi istantaneamente un volume in un point-in-time precedente. Poiché tutti i dati sul volume vengono ripristinati, VBSR potrebbe non essere appropriato per tutti i casi di utilizzo. Ad esempio, se un intero database, inclusi file di dati, log di ripristino e log di archivio, viene memorizzato in un singolo volume e questo volume viene ripristinato con VBSR, i dati vengono persi perché i log di archivio e i dati di ripristino più recenti vengono scartati.

VBSR non è necessario per il ripristino. Molti database possono essere ripristinati utilizzando SFSR (Single-file SnapRestore) basato su file o semplicemente copiando i file dalla snapshot nel file system attivo.

VBSR è preferibile quando un database è molto grande o quando deve essere recuperato il più rapidamente possibile, e l'uso di VBSR richiede l'isolamento dei file di dati. In un ambiente NFS, i file di dati di un dato database devono essere archiviati in volumi dedicati che non sono contaminati da alcun altro tipo di file. In un ambiente SAN, i file di dati devono essere memorizzati in LUN dedicate su volumi dedicati. Se viene utilizzato un volume manager (incluso Oracle Automatic Storage Management [ASM]), il gruppo di dischi deve essere dedicato anche ai file di dati.

L'isolamento dei file di dati in questo modo consente loro di tornare a uno stato precedente senza danneggiare altri file system.

Riserva di Snapshot

Per ogni volume con i dati Oracle in un ambiente SAN, il `percent-snapshot-space` Dovrebbe essere impostato su zero perché non è utile riservare spazio per uno snapshot in un ambiente LUN. Se la riserva frazionaria è impostata su 100, uno snapshot di un volume con LUN richiede spazio libero sufficiente nel volume, esclusa la riserva snapshot, per assorbire il 100% di turnover di tutti i dati. Se la riserva frazionaria è impostata su un valore inferiore, è necessaria una quantità di spazio libero corrispondente inferiore, ma esclude sempre la riserva istantanea. Ciò significa che viene sprecato lo spazio di riserva di Snapshot in un ambiente LUN.

In un ambiente NFS, esistono due opzioni:

- Impostare `percent-snapshot-space` in base al consumo di spazio snapshot previsto.
- Impostare `percent-snapshot-space` a zero e gestire collettivamente il consumo di spazio attivo e snapshot.

Con la prima opzione, `percent-snapshot-space` è impostato su un valore diverso da zero, in genere intorno al 20%. Questo spazio viene quindi nascosto all'utente. Tuttavia, questo valore non crea un limite di utilizzo. Se un database con una prenotazione del 20% registra un fatturato del 30%, lo spazio snapshot può crescere oltre i limiti della riserva del 20% e occupare spazio non riservato.

Il vantaggio principale dell'impostazione di una riserva a un valore come 20% è verificare che una parte di spazio sia sempre disponibile per gli snapshot. Ad esempio, un volume da 1TB TB con una riserva del 20% consentirebbe all'amministratore di database (DBA) di memorizzare 800GB TB di dati. Questa configurazione garantisce almeno 200GB GB di spazio per il consumo di snapshot.

Quando `percent-snapshot-space` è impostato su zero, tutto lo spazio nel volume è disponibile per l'utente finale, il che garantisce una migliore visibilità. Un DBA deve capire che, se rileva un volume di 1TB GB che sfrutta le snapshot, questo 1TB GB di spazio viene condiviso tra i dati attivi e il turnover di Snapshot.

Non esiste una chiara preferenza tra l'opzione 1 e l'opzione 2 tra gli utenti finali.

ONTAP e snapshot di terze parti

Oracle Doc ID 604683,1 illustra i requisiti per il supporto di snapshot di terze parti e le varie opzioni disponibili

per le operazioni di backup e ripristino.

Il fornitore di terze parti deve garantire che le istantanee dell'azienda siano conformi ai seguenti requisiti:

- Gli snapshot devono integrarsi con le operazioni di ripristino e ripristino consigliate da Oracle.
- Gli snapshot devono essere coerenti con il crash del database nel punto dello snapshot.
- L'ordine di scrittura viene mantenuto per ogni file all'interno di uno snapshot.

I prodotti di gestione ONTAP e NetApp di Oracle sono conformi a questi requisiti.

SnapRestore

La tecnologia NetApp SnapRestore offre il ripristino rapido dei dati in ONTAP a partire da una snapshot.

Quando un set di dati critico non è disponibile, le operazioni di business critiche non sono attive. I nastri possono interrompersi e persino i ripristini da backup basati su disco possono essere lenti da trasferire sulla rete. SnapRestore consente di evitare questi problemi grazie al ripristino quasi istantaneo dei set di dati. Anche i database di diversi petabyte possono essere ripristinati completamente con pochi minuti di lavoro.

Esistono due forme di SnapRestore: Basata su file/LUN e basata su volume.

- Singoli file o LUN possono essere ripristinati in pochi secondi, sia in una LUN da 2TB GB che in un file da 4KB GB.
- Il container di file o LUN può essere ripristinato in pochi secondi, siano essi 10GB o 100TB TB di dati.

Un "contenitore di file o LUN" generalmente si riferisce a un volume FlexVol. Ad esempio, potresti avere 10 LUN che costituiscono un gruppo di dischi LVM in un singolo volume, oppure un volume potrebbe archiviare le home directory NFS di 1000 utenti. Invece di eseguire un'operazione di ripristino per ogni singolo file o LUN, è possibile ripristinare l'intero volume come un'unica operazione. Questo processo funziona anche con container scale-out che includono volumi multipli, come FlexGroup o un gruppo di coerenza ONTAP.

Il motivo per cui SnapRestore funziona in modo così rapido ed efficiente è dovuto alla natura di uno snapshot, che è essenzialmente una vista parallela di sola lettura del contenuto di un volume in uno specifico momento. I blocchi attivi sono i blocchi reali che è possibile modificare, mentre lo snapshot è una vista di sola lettura dello stato dei blocchi che costituiscono i file e le LUN al momento della creazione dello snapshot.

ONTAP consente solo l'accesso in sola lettura ai dati snapshot, ma i dati possono essere riattivati con SnapRestore. Lo snapshot viene riabilitato come visualizzazione lettura-scrittura dei dati, riportando i dati allo stato precedente. SnapRestore può operare a livello di volume o di file. La tecnologia è essenzialmente la stessa con alcune differenze minori nel comportamento.

SnapRestore volume

La SnapRestore basata su volume riporta l'intero volume di dati a uno stato precedente. Questa operazione non richiede lo spostamento dei dati, il che significa che il processo di ripristino è essenzialmente istantaneo, sebbene l'elaborazione delle operazioni API o CLI possa richiedere alcuni secondi. Il ripristino di 1GB TB di dati non è più complicato o richiede molto tempo rispetto al ripristino di 1PB TB di dati. Questa funzionalità è il motivo principale per cui molti clienti aziendali migrano ai sistemi storage ONTAP. Offre un RTO misurato in secondi anche per i set di dati più grandi.

Uno svantaggio di SnapRestore basato su volumi è causato dal fatto che le modifiche all'interno di un volume sono cumulative nel tempo. Pertanto, ogni snapshot e i dati del file attivo dipendono dalle modifiche che hanno

portato a quel punto. Ripristinare uno stato precedente di un volume significa ignorare tutte le modifiche successive apportate ai dati. Ciò che è meno ovvio, tuttavia, è che questo include gli snapshot creati successivamente. Ciò non è sempre desiderabile.

Ad esempio, uno SLA di conservazione dei dati può specificare 30 giorni di backup notturni. Il ripristino di un set di dati in uno snapshot creato cinque giorni fa con Volume SnapRestore scaricherebbe tutti gli snapshot creati nei cinque giorni precedenti, violando lo SLA.

Sono disponibili diverse opzioni per risolvere questo limite:

1. I dati possono essere copiati da una snapshot precedente, invece di eseguire un SnapRestore dell'intero volume. Questo metodo funziona meglio con set di dati più piccoli.
2. È possibile clonare una snapshot invece di ripristinarla. Il limite a questo approccio è che lo snapshot di origine è una dipendenza del clone. Pertanto, non può essere eliminato a meno che il clone non venga anch'esso eliminato o diviso in un volume indipendente.
3. Utilizzo di SnapRestore basati su file.

File SnapRestore (Stato file)

SnapRestore basato su file è un processo di ripristino più granulare e basato su snapshot. Invece di ripristinare lo stato di un intero volume, viene ripristinato lo stato di un singolo file o LUN. Non è necessario eliminare gli snapshot, né questa operazione crea alcuna dipendenza da uno snapshot precedente. Il file o LUN diventa immediatamente disponibile nel volume attivo.

Durante il ripristino di SnapRestore di un file o LUN non è necessario alcuno spostamento dei dati. Tuttavia, alcuni aggiornamenti dei metadati interni sono necessari per riflettere il fatto che i blocchi sottostanti in un file o LUN ora esistono sia in una snapshot che nel volume attivo. Non dovrebbe avere alcun effetto sulle prestazioni, ma questo processo blocca la creazione di snapshot fino al completamento. La velocità di elaborazione è di circa 5Gbps MB (18TB MB/ora) in base alla dimensione totale dei file ripristinati.

Backup in linea

Per proteggere e ripristinare un database Oracle in modalità backup sono richiesti due set di dati. Si noti che questa non è l'unica opzione di backup di Oracle, ma è la più comune.

- Un'istantanea dei file di dati in modalità di backup
- I registri di archivio creati mentre i file di dati erano in modalità backup

Se è richiesto il recupero completo, comprese tutte le transazioni impegnate, è necessario un terzo elemento:

- Una serie di registri di ripristino correnti

Esistono diversi modi per eseguire il ripristino di un backup online. Molti clienti ripristinano le snapshot utilizzando l'interfaccia CLI di ONTAP e quindi Oracle RMAN o sqlplus per completare il ripristino. Ciò è particolarmente comune negli ambienti di produzione di grandi dimensioni, in cui la probabilità e la frequenza dei ripristini dei database sono estremamente ridotte e qualsiasi procedura di ripristino viene gestita da un DBA esperto. Per un'automazione completa, soluzioni come NetApp SnapCenter includono un plug-in Oracle con interfacce sia a riga di comando che grafiche.

Alcuni clienti su larga scala hanno adottato un approccio più semplice configurando script di base sugli host per impostare i database in modalità di backup in un momento specifico in preparazione a uno snapshot pianificato. Ad esempio, pianificare il comando `alter database begin backup` alle 23:58, `alter`

`database end backup` alle 00:02, quindi programmare le snapshot direttamente sul sistema storage a mezzanotte. Il risultato è una strategia di backup semplice e altamente scalabile che non richiede licenze o software esterni.

Layout dei dati

Il layout più semplice consiste nell'isolare i file di dati in uno o più volumi dedicati. Non devono essere contaminati da alcun altro tipo di file. In questo modo si garantisce che i volumi dei file dati possano essere ripristinati rapidamente tramite un'operazione SnapRestore senza distruggere un log di ripristino, controlfile o un log di archivio importante.

LE SAN hanno requisiti simili per l'isolamento dei file dati all'interno di volumi dedicati. Con un sistema operativo come Microsoft Windows, un singolo volume potrebbe contenere più LUN di file dati, ciascuno con un file system NTFS. Con altri sistemi operativi, in genere esiste un volume manager logico. Ad esempio, con Oracle ASM, l'opzione più semplice sarebbe limitare i LUN di un gruppo di dischi ASM a un singolo volume che può essere sottoposto a backup e ripristinato come unità. Se per motivi di gestione delle performance o della capacità sono necessari volumi aggiuntivi, la creazione di un gruppo di dischi aggiuntivo sul nuovo volume semplifica la gestione.

Se vengono seguite queste linee guida, le snapshot possono essere pianificate direttamente sul sistema di storage, senza che sia necessario eseguire uno snapshot del gruppo di coerenza. Il motivo è che i backup Oracle non richiedono il backup dei file di dati contemporaneamente. La procedura di backup online è stata progettata per consentire ai file di dati di continuare ad essere aggiornati, poiché vengono lentamente trasmessi su nastro nel corso delle ore.

Una complicazione si verifica in situazioni come l'utilizzo di un gruppo di dischi ASM distribuito tra i volumi. In questi casi, è necessario eseguire uno snapshot cg per assicurarsi che i metadati ASM siano coerenti in tutti i volumi costituenti.

Attenzione: verificare che l'ASM `spfile` e. `passwd` i file non si trovano nel gruppo di dischi che ospita i file di dati. Ciò interferisce con la capacità di ripristinare selettivamente i dati e solo i file di dati.

Procedura di ripristino locale: NFS

Questa procedura può essere gestita manualmente o tramite un'applicazione come SnapCenter. La procedura di base è la seguente:

1. Arrestare il database.
2. Recuperare i volumi di file dati nello snapshot immediatamente prima del punto di ripristino desiderato.
3. Riprodurre i log di archivio nel punto desiderato.
4. Se si desidera completare il ripristino, riprodurre i registri di ripristino correnti.

Questa procedura presuppone che i log di archivio desiderati siano ancora presenti nel file system attivo. In caso contrario, è necessario ripristinare i log di archivio oppure è possibile indirizzare rman/sqlplus ai dati nella directory snapshot.

Inoltre, per i database di dimensioni inferiori, i file di dati possono essere recuperati da un utente finale direttamente da `.snapshot` directory senza l'assistenza di tool di automazione o amministratori dello storage per eseguire una `snaprestore` comando.

Procedura di ripristino locale: SAN

Questa procedura può essere gestita manualmente o tramite un'applicazione come SnapCenter. La procedura di base è la seguente:

1. Arrestare il database.
2. Chiudere i gruppi di dischi che ospitano i file di dati. La procedura varia a seconda del volume manager logico scelto. Con ASM, il processo richiede lo smontaggio del gruppo di dischi. Con Linux, i file system devono essere smontati e i volumi logici e i gruppi di volumi devono essere disattivati. L'obiettivo è quello di interrompere tutti gli aggiornamenti del gruppo di volumi di destinazione da ripristinare.
3. Ripristinare i gruppi di dischi del file dati nello snapshot immediatamente prima del punto di ripristino desiderato.
4. Riattivare i gruppi di dischi appena ripristinati.
5. Riprodurre i log di archivio nel punto desiderato.
6. Se si desidera eseguire il ripristino completo, riprodurre tutti i registri di ripristino.

Questa procedura presuppone che i log di archivio desiderati siano ancora presenti nel file system attivo. In caso contrario, è necessario ripristinare i registri di archivio portando i LUN del registro di archivio offline ed eseguendo un ripristino. Questo è anche un esempio in cui è utile dividere i log di archivio in volumi dedicati. Se i log dell'archivio condividono un gruppo di volumi con log di ripristino, i log di ripristino devono essere copiati in un altro punto prima di ripristinare il set complessivo di LUN. Questa fase impedisce la perdita di tali transazioni finali registrate.

Backup ottimizzati per le snapshot di storage

Il backup e il ripristino basati su Snapshot sono diventati ancora più semplici quando è stato rilasciato Oracle 12c perché non è necessario collocare un database in modalità hot backup. Il risultato è la possibilità di pianificare backup basati su snapshot direttamente in un sistema storage, preservando comunque la capacità di eseguire ripristini completi o point-in-time.

Sebbene la procedura di ripristino con backup a caldo sia più familiare per gli amministratori di database, da molto tempo è stato possibile utilizzare istantanee che non sono state create mentre il database era in modalità di backup a caldo. Per rendere il database coerente, sono stati necessari ulteriori passaggi manuali con Oracle 10g e 11g durante il ripristino. Con Oracle 12c, `sqlplus` e. `rman` contenere la logica aggiuntiva per riprodurre i log di archivio sui backup dei file dati che non erano in modalità hot backup.

Come indicato in precedenza, il ripristino di un backup a caldo basato su snapshot richiede due set di dati:

- Un'istantanea dei file di dati creati in modalità backup
- I log di archivio generati mentre i file di dati erano in modalità hot backup

Durante il ripristino, il database legge i metadati dai file di dati per selezionare i log di archivio richiesti per il ripristino.

Per ottenere gli stessi risultati, il recovery ottimizzato per le snapshot di storage richiede set di dati leggermente diversi:

- Un'istantanea dei file di dati, più un metodo per identificare l'ora in cui è stata creata l'istantanea
- Archiviare i log dall'ora del checkpoint del file dati più recente all'ora esatta dello snapshot

Durante il ripristino, il database legge i metadati dai file di dati per identificare il registro di archivio più recente richiesto. È possibile eseguire il ripristino completo o point-in-time. Quando si esegue un ripristino point-in-time, è fondamentale conoscere l'ora dello snapshot dei file di dati. Il punto di ripristino specificato deve essere successivo all'ora di creazione degli snapshot. NetApp consiglia di aggiungere almeno alcuni minuti all'ora

dello snapshot per tenere conto della variazione dell'orologio.

Per informazioni dettagliate, vedere la documentazione di Oracle sull'argomento "Recovery Using Storage Snapshot Optimization" disponibile in varie versioni della documentazione di Oracle 12c. Inoltre, consultare l'ID documento Oracle Doc ID 604683,1 relativo al supporto per le istantanee di terze parti di Oracle.

Layout dei dati

Il layout più semplice consiste nell'isolare i file di dati in uno o più volumi dedicati. Non devono essere contaminati da alcun altro tipo di file. In questo modo si garantisce che i volumi dei file dati possano essere ripristinati rapidamente con un'operazione SnapRestore senza distruggere un log di ripristino, controlfile o un log di archivio importante.

LE SAN hanno requisiti simili per l'isolamento dei file dati all'interno di volumi dedicati. Con un sistema operativo come Microsoft Windows, un singolo volume potrebbe contenere più LUN di file dati, ciascuno con un file system NTFS. Con altri sistemi operativi, esiste in genere anche un volume manager logico. Ad esempio, con Oracle ASM, l'opzione più semplice sarebbe quella di limitare i gruppi di dischi a un singolo volume di cui è possibile eseguire il backup e il ripristino come unità. Se per motivi di gestione delle performance o della capacità sono necessari volumi aggiuntivi, la creazione di un gruppo di dischi aggiuntivo sul nuovo volume semplifica la gestione.

Se si seguono queste linee guida, gli snapshot possono essere pianificati direttamente su ONTAP senza che sia necessario eseguire uno snapshot del gruppo di coerenza. Il motivo è che i backup ottimizzati per le istantanee non richiedono che venga eseguito contemporaneamente il backup dei file di dati.

Una complicazione si verifica in situazioni come un gruppo di dischi ASM distribuito tra i volumi. In questi casi, è necessario eseguire uno snapshot cg per assicurarsi che i metadati ASM siano coerenti in tutti i volumi costituenti.

[Note]verificare che i file ASM spfile e passwd non siano nel gruppo di dischi che ospita i file di dati. Ciò interferisce con la capacità di ripristinare selettivamente i dati e solo i file di dati.

Procedura di ripristino locale: NFS

Questa procedura può essere gestita manualmente o tramite un'applicazione come SnapCenter. La procedura di base è la seguente:

1. Arrestare il database.
2. Recuperare i volumi di file dati nello snapshot immediatamente prima del punto di ripristino desiderato.
3. Riprodurre i log di archivio nel punto desiderato.

Questa procedura presuppone che i log di archivio desiderati siano ancora presenti nel file system attivo. In caso contrario, è necessario ripristinare i registri di archivio, o. `rman` oppure `sqlplus` può essere indirizzato ai dati in `.snapshot` directory.

Inoltre, per i database di dimensioni inferiori, i file di dati possono essere recuperati da un utente finale direttamente da `.snapshot` Senza l'assistenza di tool di automazione o di un amministratore dello storage per eseguire un comando SnapRestore.

Procedura di ripristino locale: SAN

Questa procedura può essere gestita manualmente o tramite un'applicazione come SnapCenter. La procedura di base è la seguente:

1. Arrestare il database.
2. Chiudere i gruppi di dischi che ospitano i file di dati. La procedura varia a seconda del volume manager logico scelto. Con ASM, il processo richiede lo smontaggio del gruppo di dischi. Con Linux, i file system devono essere smontati e i volumi logici e i gruppi di volumi sono disattivati. L'obiettivo è quello di interrompere tutti gli aggiornamenti del gruppo di volumi di destinazione da ripristinare.
3. Ripristinare i gruppi di dischi del file dati nello snapshot immediatamente prima del punto di ripristino desiderato.
4. Riattivare i gruppi di dischi appena ripristinati.
5. Riprodurre i log di archivio nel punto desiderato.

Questa procedura presuppone che i log di archivio desiderati siano ancora presenti nel file system attivo. In caso contrario, è necessario ripristinare i registri di archivio portando i LUN del registro di archivio offline ed eseguendo un ripristino. Questo è anche un esempio in cui è utile dividere i log di archivio in volumi dedicati. Se i log dell'archivio condividono un gruppo di volumi con i log di ripristino, i log di ripristino devono essere copiati in un altro punto prima del ripristino del set complessivo di LUN, per evitare di perdere le transazioni finali registrate.

Esempio di recupero completo

Si supponga che i file di dati siano stati corrotti o distrutti e che sia necessario un ripristino completo. La procedura da seguire è la seguente:

```
[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>
```

Esempio di recupero point-in-time

L'intera procedura di ripristino è un singolo comando: `recover automatic`.

Se è necessario un ripristino point-in-time, l'indicatore data e ora degli snapshot deve essere noto e può essere identificato come segue:

```
Cluster01::> snapshot show -vserver vserver1 -volume NTAP_oradata -fields
create-time
vserver   volume        snapshot   create-time
--------  ------------  ---------  ------------------------
vserver1  NTAP_oradata  my-backup  Thu Mar 09 10:10:06 2017
```

L'ora di creazione dell'istantanea è indicata come marzo 9th e 10:10:06. Per essere sicuri, viene aggiunto un minuto all'ora dell'istantanea:

```
[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';
```

Il ripristino viene avviato. È stato specificato un tempo di snapshot di 10:11:00, un minuto dopo il tempo registrato per tenere conto della possibile varianza dell'orologio e un tempo di recupero target di 10:44. Successivamente, sqlplus richiede i registri di archivio necessari per raggiungere il tempo di ripristino desiderato di 10:44.

```
ORA-00279: change 551760 generated at 03/09/2017 05:06:07 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_31_930813377.dbf
ORA-00280: change 551760 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 552566 generated at 03/09/2017 05:08:09 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_32_930813377.dbf
ORA-00280: change 552566 for thread 1 is in sequence #32
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 553045 generated at 03/09/2017 05:10:12 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_33_930813377.dbf
ORA-00280: change 553045 for thread 1 is in sequence #33
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 753229 generated at 03/09/2017 05:15:58 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_34_930813377.dbf
ORA-00280: change 753229 for thread 1 is in sequence #34
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
Log applied.
Media recovery complete.
SQL> alter database open resetlogs;
Database altered.
SQL>
```

Completare il ripristino di un database utilizzando gli snapshot utilizzando `recover automatic` command non richiede licenze specifiche, ma utilizza un ripristino point-in-time `snapshot time` Richiede la licenza Oracle Advanced Compression.

Tool di gestione e automazione del database

Il valore primario di ONTAP in un ambiente di database Oracle deriva dalle tecnologie principali di ONTAP, come copie Snapshot istantanee, semplice replica SnapMirror e creazione efficiente dei volumi FlexClone.

In alcuni casi, una semplice configurazione di queste funzionalità chiave direttamente su ONTAP soddisfa i requisiti, ma esigenze più complesse richiedono un livello di orchestrazione.

SnapCenter

SnapCenter è il prodotto di punta della protezione dei dati di NetApp. A un livello molto basso, è simile ai prodotti SnapManager in termini di modalità di esecuzione dei backup del database, ma è stato creato da zero per fornire un singolo pannello di controllo per la gestione della protezione dati sui sistemi di storage NetApp.

SnapCenter include le funzioni di base come backup e ripristini basati su snapshot, la replica SnapMirror e SnapVault e altre funzionalità necessarie per operare su larga scala per le grandi imprese. Queste funzionalità

avanzate includono una funzionalità estesa di controllo degli accessi in base al ruolo (RBAC), API RESTful per l'integrazione con prodotti di orchestrazione di terze parti, gestione centrale senza interruzioni dei plug-in SnapCenter sugli host di database e un'interfaccia utente progettata per ambienti cloud-scale.

RIPOSO

ONTAP contiene anche un ricco set di API RESTful. Questo consente a 3rd vendor di creare data Protection e altre applicazioni di gestione con una profonda integrazione con ONTAP. Inoltre, l'API RESTful è facile da utilizzare da parte dei clienti che desiderano creare i propri flussi di lavoro e utility di automazione.

Disaster recovery Oracle

Panoramica

Il disaster recovery si riferisce al ripristino dei servizi dati dopo un evento catastrofico, come un incendio che distrugge un sistema storage o persino un'intera sede.

Questa documentazione sostituisce i report tecnici precedentemente pubblicati *TR-4591: Oracle Data Protection* e *TR-4592: Oracle on MetroCluster.*

Il disaster recovery può essere eseguito mediante una semplice replica dei dati tramite SnapMirror, naturalmente, con molti clienti che aggiornano le repliche con mirroring ogni ora.

Per la maggior parte dei clienti, il disaster recovery non richiede solo una copia remota dei dati, ma anche la capacità di sfruttarli in maniera rapida. NetApp offre due tecnologie che soddisfano questa esigenza: MetroCluster e SnapMirror Active Sync

MetroCluster fa riferimento a ONTAP in una configurazione hardware che include storage con mirroring sincrono di basso livello e numerose funzionalità aggiuntive. Le soluzioni integrate come MetroCluster semplificano le complesse e scalabili infrastrutture di database, applicazioni e virtualizzazione. Sostituisce diversi prodotti e strategie di protezione dati esterni con un unico semplice storage array centrale. Fornisce inoltre backup, recovery, disaster recovery e alta disponibilità (ha) integrati in un singolo sistema storage in cluster.

La sincronizzazione attiva di SnapMirror (SM-AS) si basa sulla sincronizzazione sincrona di SnapMirror. Con MetroCluster, ogni controller ONTAP è responsabile della replica dei dati dell'unità in una posizione remota. Con la sincronizzazione attiva di SnapMirror, avrai essenzialmente due sistemi ONTAP diversi che mantengono copie indipendenti dei dati LUN, ma cooperano per presentare una singola istanza di tale LUN. Dal punto di vista dell'host, si tratta di una singola entità LUN.

Confronto SM-AS e MCC

SM-AS e MetroCluster sono simili per quanto riguarda le funzionalità generali, ma esistono importanti differenze nel modo in cui è stata implementata la replica RPO=0 e nel modo in cui viene gestita. Anche se è possibile utilizzare la modalità asincrona e sincrona di SnapMirror come parte di un piano di disaster recovery, non sono progettate come tecnologie di replica ha.

- Una configurazione MetroCluster è più simile a un cluster integrato con nodi distribuiti tra i siti. SM-AS si comporta come due cluster altrimenti indipendenti che stanno cooperando nel servire un RPO selezionato=0 LUN replicati in modo sincrono.
- I dati in una configurazione MetroCluster sono accessibili solo da un determinato sito alla volta. Una seconda copia dei dati è presente sul sito opposto, ma i dati sono passivi. Non è possibile accedervi senza un failover del sistema storage.

- MetroCluster e SM-as eseguono il mirroring a diversi livelli. Il mirroring MetroCluster viene eseguito al livello RAID. I dati di basso livello sono memorizzati in un formato di mirroring utilizzando SyncMirror. L'utilizzo del mirroring è praticamente invisibile ai livelli di LUN, volume e protocollo.
- Al contrario, il mirroring SM-AS avviene a livello di protocollo. I due cluster sono complessivamente cluster indipendenti. Una volta sincronizzate le due copie di dati, i due cluster devono solo eseguire il mirroring delle scritture. Quando si verifica una scrittura su un cluster, questa viene replicata nell'altro cluster. La scrittura viene riconosciuta all'host solo quando la scrittura è stata completata su entrambi i siti. A parte questo comportamento di suddivisione del protocollo, i due cluster sono altrimenti normali cluster ONTAP.
- Il ruolo principale di MetroCluster è la replica su larga scala. Puoi replicare un intero array con RPO=0 e RTO prossimo allo zero. Questo semplifica il processo di failover perché esiste un solo "problema" da eseguire e consente una scalabilità perfetta in termini di capacità e IOPS.
- Un caso d'utilizzo chiave per SM-AS è la replica granulare. A volte non vuoi replicare tutti i dati come una singola unità oppure devi eseguire il failover selettivo su alcuni carichi di lavoro.
- Un altro caso d'utilizzo chiave per SM-AS è per operazioni Active-Active, dove desideri che siano disponibili copie dei dati completamente utilizzabili su due cluster diversi situati in due posizioni diverse con caratteristiche di performance identiche e, se desiderato, non richiedere l'estensione della SAN tra i siti. Le applicazioni possono essere già in esecuzione su entrambi i siti, riducendo così l'RTO complessivo durante le operazioni di failover.

MetroCluster

Disaster recovery con MetroCluster

MetroCluster è una funzionalità di ONTAP in grado di proteggere i database Oracle con RPO=0 mirroring sincrono tra i siti, per poi scalare in verticale e supportare centinaia di database su un singolo sistema MetroCluster.

È anche semplice da usare. L'utilizzo di MetroCluster non aggiunge o modifica necessariamente i migliori percorsi per la gestione di applicazioni e database aziendali.

Le normali Best practice vengono comunque applicate e se le tue esigenze richiedono solo RPO=0:1 di data Protection, allora MetroCluster ne soddisfa l'esigenza. Tuttavia, la maggior parte dei clienti utilizza MetroCluster non solo per la protezione dei dati con RPO=0, ma anche per migliorare l'RTO in scenari di disastro, oltre a fornire un failover trasparente come parte delle attività di manutenzione del sito.

Architettura fisica

Per comprendere il funzionamento dei database Oracle in un ambiente MetroCluster è necessario spiegare la progettazione fisica di un sistema MetroCluster.

> Questa documentazione sostituisce il report tecnico precedentemente pubblicato *TR-4592: Oracle su MetroCluster.*

MetroCluster è disponibile in 3 diverse configurazioni

- Coppie HA con connettività IP
- Coppie HA con connettività FC
- Controller singolo con connettività FC

Il termine connettività si riferisce alla connessione cluster utilizzata per la replica tra siti. Non si riferisce ai protocolli host. Tutti i protocolli lato host sono supportati come di consueto in una configurazione MetroCluster indipendentemente dal tipo di connessione utilizzata per la comunicazione tra cluster.

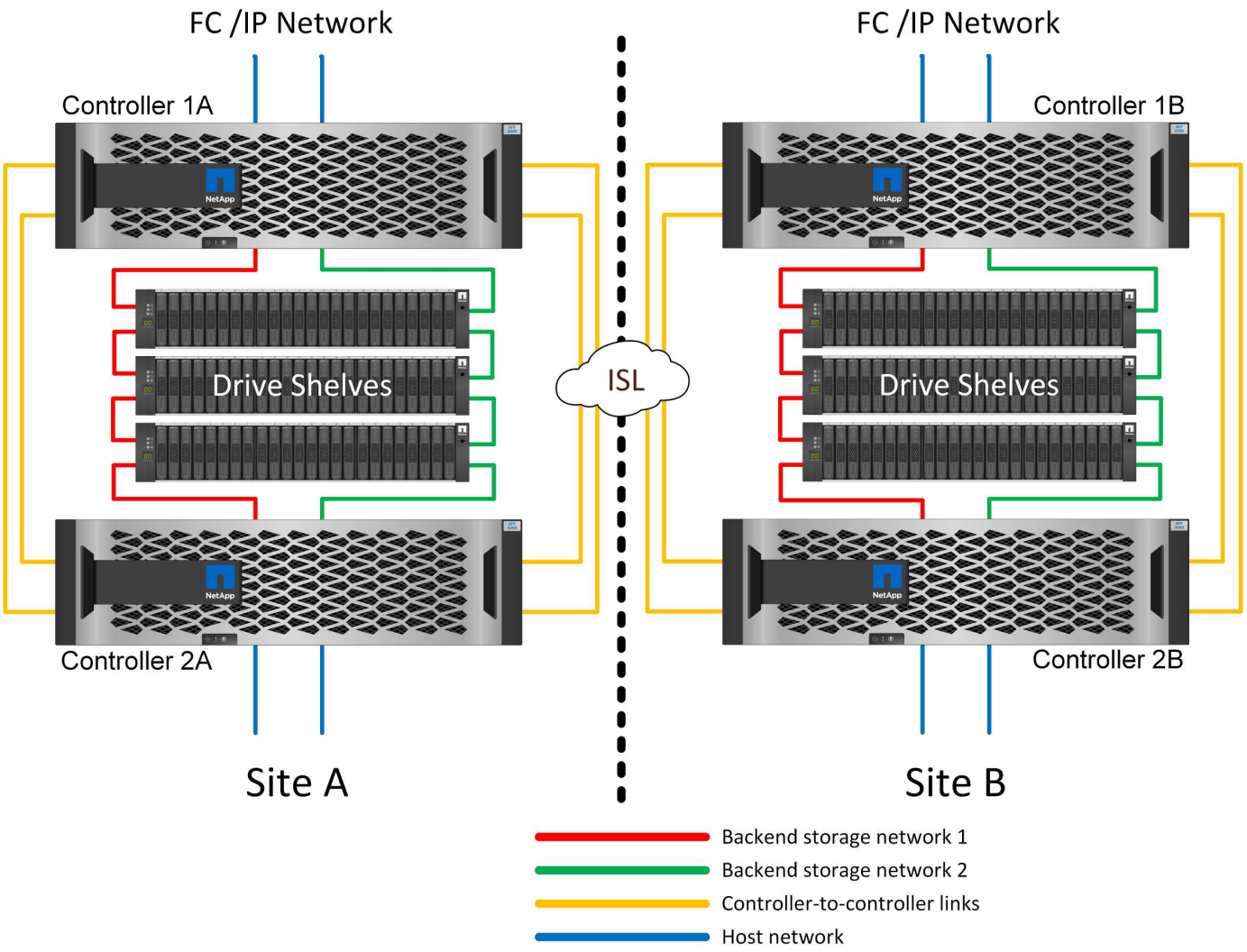
IP MetroCluster

La configurazione MetroCluster IP ha-Pair utilizza due o quattro nodi per sito. Questa opzione di configurazione aumenta la complessità e i costi rispetto all'opzione a due nodi, ma offre un vantaggio importante: La ridondanza intrasite. Un semplice errore del controller non richiede l'accesso ai dati nella WAN. L'accesso ai dati rimane locale attraverso il controller locale alternativo.

La maggior parte dei clienti sceglie la connettività IP perché i requisiti dell'infrastruttura sono più semplici. In passato, la connettività cross-site ad alta velocità era generalmente più semplice da fornire utilizzando gli switch FC e in fibra scura, ma oggi i circuiti IP ad alta velocità e a bassa latenza sono più prontamente disponibili.

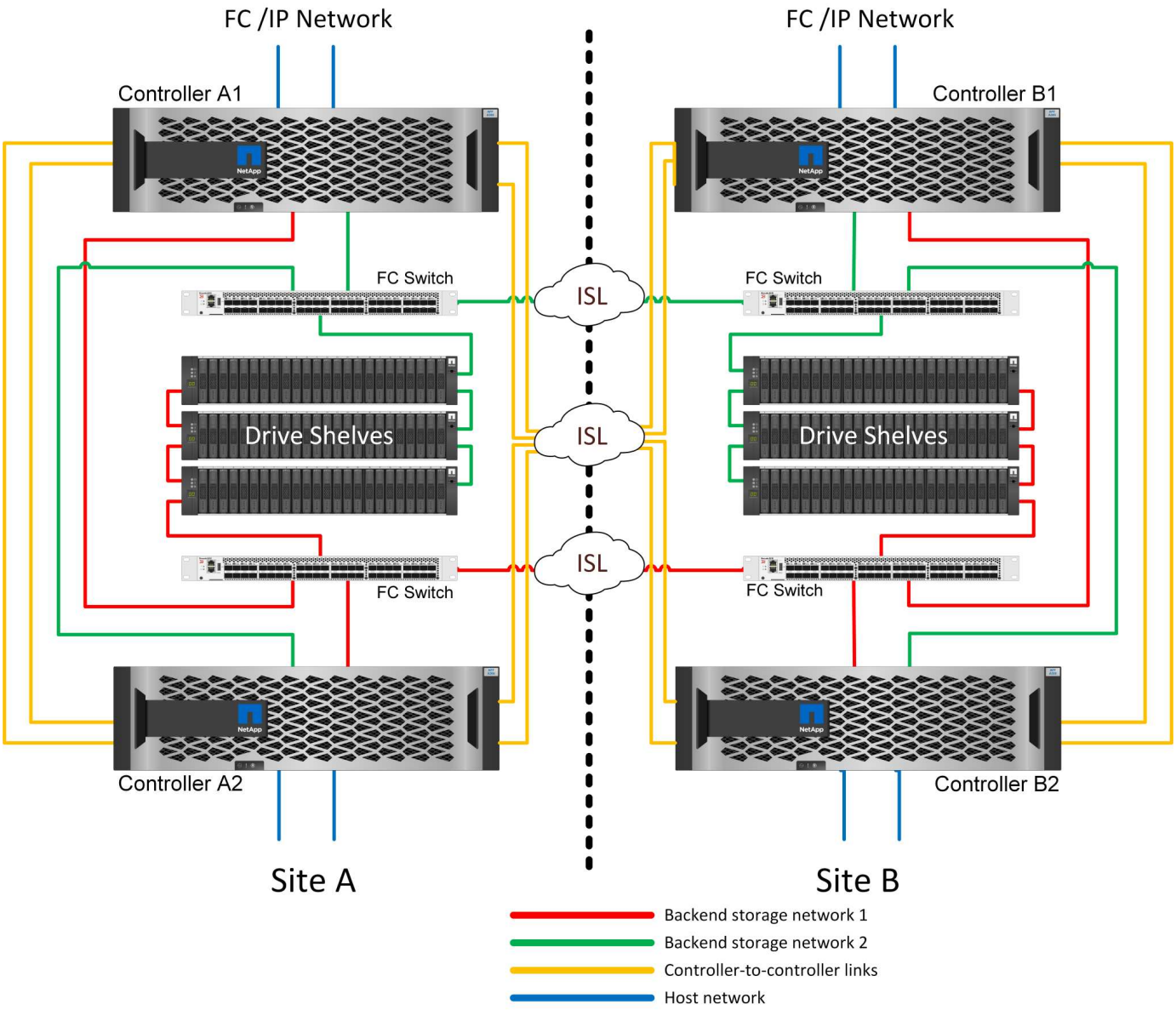
L'architettura è anche più semplice perché le uniche connessioni cross-site sono per i controller. Nei MetroClusters collegati a FC SAN, un controller scrive direttamente sulle unità del sito opposto e quindi richiede connessioni SAN, switch e bridge aggiuntivi. Al contrario, un controller in una configurazione IP scrive sulle unità opposte tramite il controller.

Per ulteriori informazioni, consultare la documentazione ufficiale di ONTAP e. "Architettura e progettazione della soluzione IP di MetroCluster".



MetroCluster HA-Pair FC SAN-Attached

La configurazione ha-Pair MetroCluster FC utilizza due o quattro nodi per sito. Questa opzione di configurazione aumenta la complessità e i costi rispetto all'opzione a due nodi, ma offre un vantaggio importante: La ridondanza intrasite. Un semplice errore del controller non richiede l'accesso ai dati nella WAN. L'accesso ai dati rimane locale attraverso il controller locale alternativo.

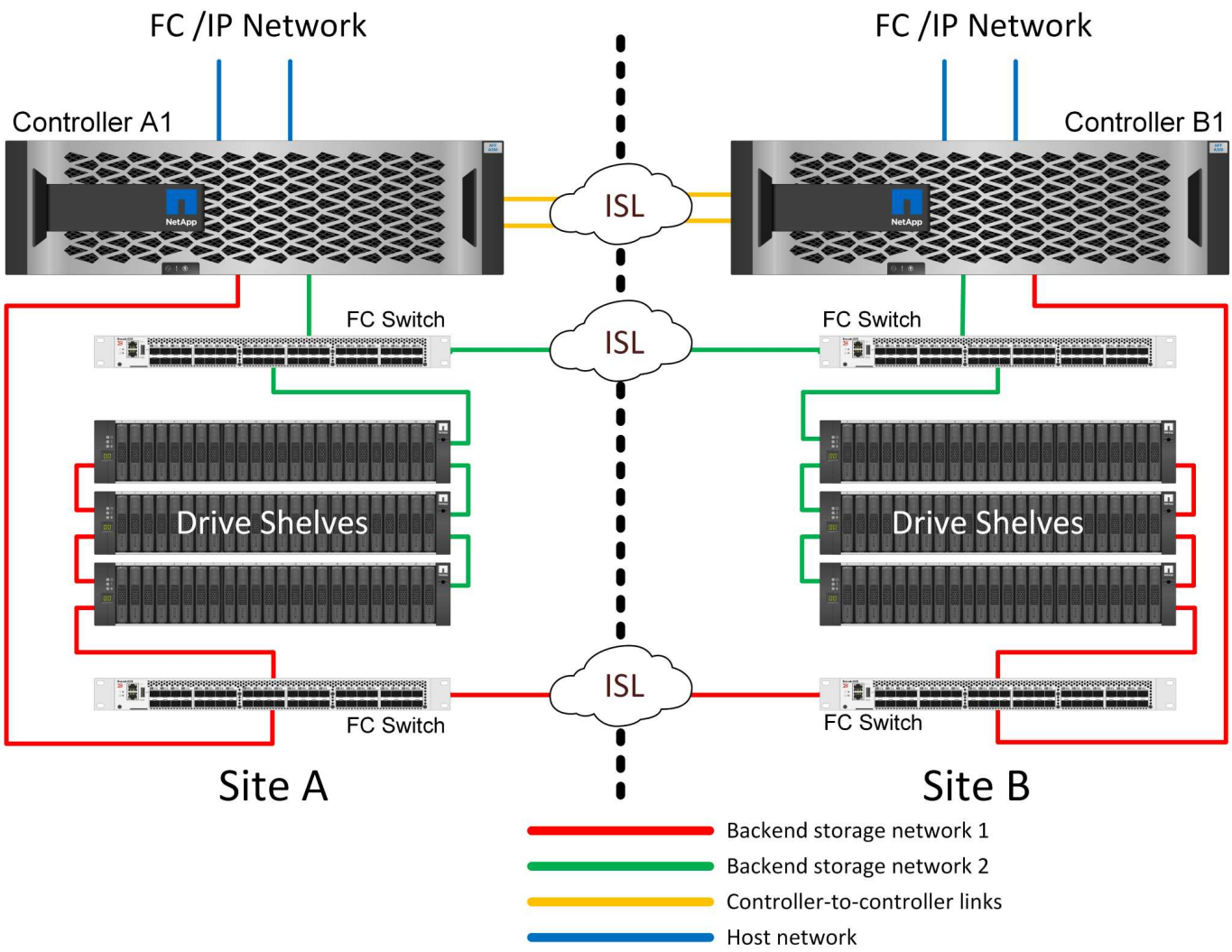


Alcune infrastrutture multisito non sono progettate per le operazioni Active-Active, ma vengono utilizzate maggiormente come sito primario e sito di disaster recovery. In questa situazione, è generalmente preferibile un'opzione ha-Pair MetroCluster per i seguenti motivi:

- Anche se un cluster MetroCluster a due nodi è un sistema ha, un guasto imprevisto di un controller o una manutenzione pianificata richiedono che i servizi dati vengano online sul sito opposto. Se la connettività di rete tra i siti non supporta la larghezza di banda richiesta, le prestazioni ne risentono. L'unica opzione sarebbe anche eseguire il failover dei vari sistemi operativi host e dei servizi associati al sito alternativo. Il cluster MetroCluster ha-Pair elimina questo problema grazie alla perdita di un controller che consente di eseguire un semplice failover all'interno dello stesso sito.
- Alcune topologie di rete non sono progettate per l'accesso tra siti, ma utilizzano sottoreti o SAN FC isolate. In questi casi, il cluster MetroCluster a due nodi non funziona più come sistema ha, perché il controller alternativo non può fornire dati ai server del sito opposto. L'opzione ha-Pair MetroCluster è necessaria per garantire ridondanza completa.
- Se un'infrastruttura a due siti viene vista come una singola infrastruttura ad alta disponibilità, la configurazione MetroCluster a due nodi è adatta. Tuttavia, se il sistema deve funzionare per un periodo di tempo prolungato dopo il guasto del sito, è preferibile una coppia ha perché continua a fornire ha all'interno di un singolo sito.

MetroCluster FC SAN-attached a due nodi

La configurazione MetroCluster a due nodi utilizza un solo nodo per sito. Questo design è più semplice rispetto all'opzione ha-Pair perché richiede meno componenti da configurare e gestire. Inoltre, ha ridotto le richieste di infrastruttura in termini di cablaggio e switch FC. Infine, riduce i costi.



L'evidente impatto di questa progettazione è che un guasto del controller su un singolo sito implica che i dati sono disponibili dal sito opposto. Questa restrizione non è necessariamente un problema. Molte aziende hanno operazioni di data center multisito con reti estese, ad alta velocità e a bassa latenza che funzionano essenzialmente come una singola infrastruttura. In questi casi, la configurazione preferita è la versione a due nodi di MetroCluster. Diversi service provider utilizzano attualmente sistemi a due nodi con scalabilità di petabyte.

Funzionalità di resilienza di MetroCluster

Non esistono single point of failure in una soluzione MetroCluster:

- Ogni controller dispone di due percorsi indipendenti verso gli shelf di dischi sul sito locale.
- Ogni controller dispone di due percorsi indipendenti verso gli shelf di dischi sul sito remoto.
- Ciascun controller dispone di due percorsi indipendenti verso i controller sul sito opposto.
- Nella configurazione ha-Pair, ogni controller ha due percorsi verso il partner locale.

Riassumendo, qualsiasi componente della configurazione può essere rimosso senza compromettere la capacità di MetroCluster di fornire dati. L'unica differenza in termini di resilienza tra le due opzioni è che la versione ha-Pair è ancora un sistema storage ha generale dopo un guasto del sito.

Architettura logica

Per comprendere il funzionamento dei database Oracle in un ambiente MetroCluster alsop è necessario spiegare alcune delle funzionalità logiche di un sistema MetroCluster.

Protezione da errori del sito: NVRAM e MetroCluster

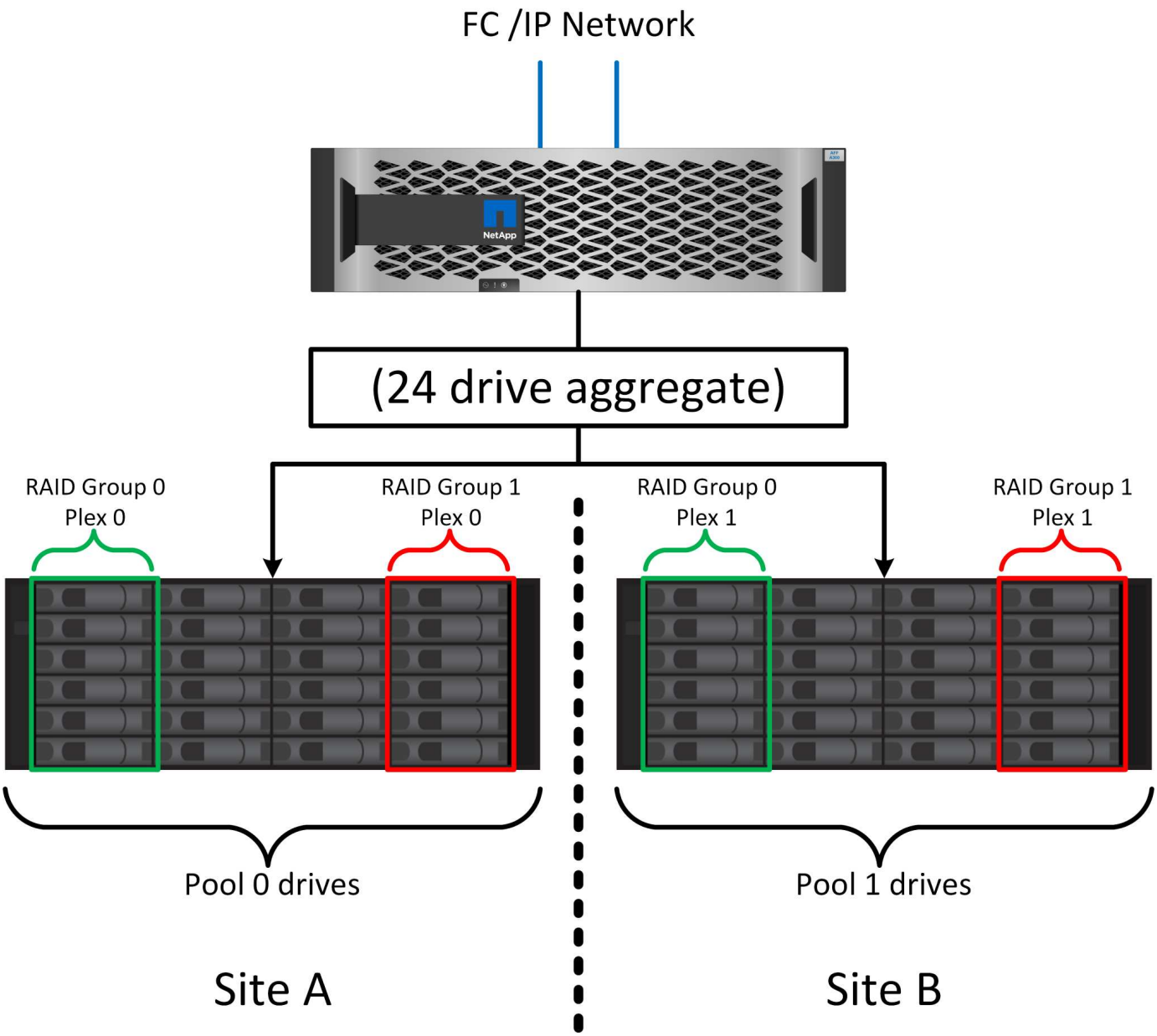
MetroCluster estende la protezione dei dati NVRAM nei seguenti modi:

- In una configurazione a due nodi, i dati NVRAM vengono replicati attraverso i collegamenti Inter-Switch (ISL) al partner remoto.
- In una configurazione ha-Pair, i dati NVRAM vengono replicati sia nel partner locale che in un partner remoto.
- Una scrittura non viene riconosciuta fino a quando non viene replicata a tutti i partner. Questa architettura protegge gli i/o in fase di trasferimento dai guasti del sito replicando i dati NVRAM a un partner remoto. Il processo non è coinvolto nella replica dei dati a livello di unità. Il controller proprietario degli aggregati si occupa della replica dei dati per iscritto a entrambi i plessi dell'aggregato, ma in caso di perdita del sito occorre comunque proteggere dalle perdite di i/o in fase di trasferimento. I dati NVRAM replicati sono utilizzati solo se un partner controller deve subentrare a un controller guasto.

Protezione dai guasti di shelf e siti: SyncMirror e plessi

SyncMirror è una tecnologia di mirroring che migliora, ma non sostituisce, RAID DP o RAID-TEC. Esegue il mirroring del contenuto di due gruppi RAID indipendenti. La configurazione logica è la seguente:

1. I dischi sono configurati in due pool in base alla posizione. Un pool è composto da tutti i dischi sul sito A, mentre il secondo è composto da tutti i dischi sul sito B.
2. Viene quindi creato un pool di storage comune, detto aggregato, in base a set di gruppi RAID con mirroring. Viene ottenuto lo stesso numero di unità per ciascun sito. Ad esempio, un aggregato SyncMirror da 20 dischi sarebbe composto da 10 dischi del sito A e 10 dischi del sito B.
3. Ogni set di unità su un dato sito viene configurato automaticamente come uno o più gruppi RAID DP o RAID-TEC completamente ridondanti, indipendentemente dall'utilizzo del mirroring. Questo utilizzo di RAID sottostante il mirroring garantisce la protezione dei dati anche dopo la perdita di un sito.



La figura precedente illustra una configurazione SyncMirror di esempio. È stato creato un aggregato di 24 dischi sul controller con 12 dischi da uno shelf allocato sul sito A e 12 dischi da uno shelf allocato sul sito B. I dischi sono stati raggruppati in due gruppi RAID con mirroring. Il gruppo RAID 0 include un plesso A 6 dischi sul sito A con mirroring su un plesso A 6 dischi sul sito B. Analogamente, il gruppo RAID 1 include un plesso A 6 dischi sul sito A con mirroring su un plesso A 6 dischi sul sito B.

Di norma, SyncMirror viene utilizzato per fornire il mirroring remoto con i sistemi MetroCluster, con una copia dei dati in ciascun sito. A volte, è stato utilizzato per fornire un livello di ridondanza extra in un unico sistema. In particolare, fornisce ridondanza a livello di shelf. Uno shelf di dischi contiene già doppi controller e alimentatori e nel complesso è poco più di una lamiera, ma in alcuni casi è consigliabile garantire una protezione extra. Ad esempio, un cliente NetApp ha implementato SyncMirror per una piattaforma mobile di analytics in tempo reale utilizzata durante i test nel settore automobilistico. Il sistema è stato separato in due rack fisici forniti con alimentatori indipendenti e sistemi UPS indipendenti.

Errore di ridondanza: NVFAIL

Come discusso in precedenza, una scrittura non viene riconosciuta fino a quando non è stata registrata nella

NVRAM locale e nella NVRAM su almeno un altro controller. Questo approccio garantisce che un guasto dell'hardware o un'interruzione di corrente non comporti la perdita dell'i/o in-flight Se si verifica un guasto nella NVRAM locale o nella connettività ad altri nodi, i dati non verranno più mirrorati.

Se la NVRAM locale riporta un errore, il nodo si arresta. Questo arresto determina il failover su un partner controller quando vengono utilizzate coppie ha. Con MetroCluster, il comportamento dipende dalla configurazione complessiva scelta, ma può portare al failover automatico della nota remota. In ogni caso, nessun dato viene perso perché il controller che subisce l'errore non ha confermato l'operazione di scrittura.

Un guasto di connettività site-to-site che blocca la replica NVRAM ai nodi remoti è una situazione più complicata. Le scritture non vengono più replicate sui nodi remoti, con la possibilità di perdita di dati in caso di errore catastrofico su un controller. Cosa più importante, il tentativo di failover su un nodo diverso in queste condizioni comporta una perdita di dati.

Il fattore di controllo è se la NVRAM è sincronizzata. Se la NVRAM è sincronizzata, il failover da nodo a nodo può procedere in tutta sicurezza senza rischio di perdita di dati. In una configurazione MetroCluster, se la NVRAM e i plessi degli aggregati sottostanti sono sincronizzati, è possibile procedere con lo switchover senza rischio di perdita di dati.

ONTAP non consente alcun failover o switchover quando i dati non sono sincronizzati, a meno che non sia forzato il failover o lo switchover. La forzatura di una modifica delle condizioni in questo modo riconosce che i dati potrebbero essere lasciati indietro nel controllore originale e che la perdita di dati è accettabile.

I database e altre applicazioni sono particolarmente vulnerabili al danneggiamento se un failover o uno switchover è forzato perché mantengono cache interne di dati su disco di dimensioni maggiori. In caso di failover o switchover forzato, le modifiche riconosciute in precedenza vengono eliminate del tutto. Il contenuto dell'array di storage torna indietro nel tempo e lo stato della cache non riflette più lo stato dei dati su disco.

Per evitare questa situazione, ONTAP consente di configurare i volumi per una protezione speciale contro i guasti della NVRAM. Quando attivato, questo meccanismo di protezione determina l'ingresso di un volume nello stato chiamato NVFAIL. Questo stato causa errori di i/o che causano un crash dell'applicazione. Questo blocco causa l'arresto delle applicazioni in modo che non utilizzino dati obsoleti. I dati non devono essere persi perché i dati delle transazioni devono essere presenti nei registri. Solitamente, gli amministratori dovranno arrestare completamente gli host prima di riportare manualmente LUN e volumi in linea. Sebbene queste fasi possano comportare un certo lavoro, questo approccio è il modo più sicuro per garantire l'integrità dei dati. Non tutti i dati richiedono questa protezione, motivo per cui il comportamento di NVFAIL può essere configurato in base al volume.

Coppie HA e MetroCluster

MetroCluster è disponibile in due configurazioni: Due nodi e coppia ha. La configurazione a due nodi si comporta come una coppia ha in relazione alla NVRAM. In caso di guasto improvviso, il nodo partner può riprodurre i dati della NVRAM per rendere i dischi coerenti e garantire che non vengano perse scritture riconosciute.

La configurazione ha-Pair replica la NVRAM anche sul nodo partner locale. Un semplice guasto al controller porta a un replay della NVRAM sul nodo partner, come nel caso di una coppia ha standalone, senza MetroCluster. In caso di improvvisa perdita completa del sito, il sito remoto dispone anche della NVRAM necessaria per rendere i dischi coerenti e iniziare a fornire i dati.

Un aspetto importante di MetroCluster è che i nodi remoti non hanno accesso ai dati partner in normali condizioni operative. Ogni sito funziona essenzialmente come un sistema indipendente che può assumere la personalità del sito opposto. Questo processo, noto come switchover, include uno switchover pianificato, in cui le operazioni del sito vengono migrate senza interruzioni nel sito opposto. Include anche le situazioni non pianificate in cui si perde un sito ed è necessario uno switchover manuale o automatico come parte del

disaster recovery.

Switchover e switchback

I termini switchover e switchback si riferiscono al processo di transizione dei volumi tra controller remoti in una configurazione MetroCluster. Questo processo si applica solo ai nodi remoti. Se viene utilizzato MetroCluster in una configurazione a quattro volumi, il failover di nodo locale utilizza il medesimo processo di takeover e giveback descritto in precedenza.

Switchover e switchback pianificati

Uno switchover o uno switchback pianificato è simile a un takeover o un giveback tra i nodi. Il processo prevede diverse fasi e potrebbe richiedere alcuni minuti, ma in realtà si tratta di una transizione graduale delle risorse di storage e di rete. Il momento in cui il trasferimento del controllo avviene molto più rapidamente del tempo richiesto per l'esecuzione del comando completo.

La differenza principale tra takeover/giveback e switchover/switchback influisce sulla connettività FC SAN. Grazie al takeover/giveback locale, un host subisce la perdita di tutti i percorsi FC nel nodo locale e si affida al proprio MPIO nativo per passare ai percorsi alternativi disponibili. Le porte non vengono ricollocate. Grazie a switchover e switchback, le porte di destinazione FC virtuali sui controller passano all'altro sito. Di fatto, smettono di esistere sulla SAN per un momento e ricompaiono su un controller alternativo.

Timeout SyncMirror

SyncMirror è una tecnologia di mirroring ONTAP che fornisce protezione dai guasti agli shelf. Quando gli shelf sono separati su una distanza, il risultato è una data Protection remota.

SyncMirror non fornisce mirroring sincrono universale. Il risultato è una maggiore disponibilità. Alcuni sistemi di archiviazione utilizzano un mirroring costante tutto o niente, talvolta chiamato modalità domino. Questa forma di mirroring è limitata nell'applicazione poiché tutte le attività di scrittura devono cessare se la connessione al sito remoto viene persa. Altrimenti, una scrittura esisterebbe in un sito ma non nell'altro. Generalmente, tali ambienti sono configurati per portare le LUN offline in caso di perdita della connettività sito-sito per più di un breve periodo (ad esempio 30 secondi).

Questo comportamento è desiderabile per un piccolo sottoinsieme di ambienti. Tuttavia, la maggior parte delle applicazioni richiede una soluzione che offra una replica sincrona garantita in normali condizioni operative, ma con la possibilità di sospendere la replica. Una perdita completa della connettività da sito a sito viene spesso considerata una situazione quasi disastrosa. Generalmente, tali ambienti vengono mantenuti online e forniscono dati fino al ripristino della connettività o alla decisione formale di arrestare l'ambiente per proteggere i dati. Un requisito per l'arresto automatico dell'applicazione solo a causa di un errore di replica remota è insolito.

SyncMirror supporta i requisiti di mirroring sincrono con la flessibilità di un timeout. Se la connettività al telecomando e/o al plex viene persa, inizia il conto alla rovescia un timer di 30 secondi. Quando il contatore raggiunge 0, l'elaborazione i/o in scrittura riprende a utilizzare i dati locali. La copia remota dei dati è utilizzabile, ma viene bloccata in tempo fino a quando non viene ripristinata la connettività. La risincronizzazione sfrutta le snapshot a livello di aggregato per riportare il sistema in modalità sincrona il più rapidamente possibile.

In particolare, in molti casi, questo tipo di replica universale in modalità domino a tutto o niente è meglio implementato a livello di applicazione. Ad esempio, Oracle DataGuard include la modalità di protezione massima, che garantisce la replica a lunga istanza in tutte le circostanze. Se il collegamento di replica non riesce per un periodo superiore a un timeout configurabile, i database vengono arrestati.

Switchover automatico senza intervento dell'utente con MetroCluster fabric-attached

Lo switchover automatico non assistito (ASOLO) è una funzione MetroCluster collegata al fabric che offre un tipo di ha cross-site. Come indicato in precedenza, MetroCluster è disponibile in due tipi: Un singolo controller su ciascun sito o una coppia ha su ciascun sito. Il vantaggio principale dell'opzione ha è che l'arresto pianificato o non pianificato del controller consente comunque a tutti gli i/o di essere locali. Il vantaggio dell'opzione a nodo singolo consiste nella riduzione di costi, complessità e infrastruttura.

Il valore primario di AUSO è migliorare le capacità ha dei sistemi MetroCluster fabric-attached. Ciascun sito esegue il monitoraggio dello stato di salute del sito opposto e, se non sono ancora presenti nodi che forniscono dati, AUDO esegue un rapido switchover. Questo approccio è particolarmente utile nelle configurazioni MetroCluster con un solo nodo per sito, perché consente di avvicinare la configurazione a una coppia ha in termini di disponibilità.

AUSO non è in grado di offrire un monitoraggio completo a livello di coppia ha. Una coppia ha può offrire una disponibilità estremamente elevata, perché include due cavi fisici ridondanti per la comunicazione diretta da nodo a nodo. Inoltre, entrambi i nodi di una coppia ha hanno accesso allo stesso set di dischi in loop ridondanti, offrendo un altro percorso a un nodo per monitorare la salute di un altro.

I cluster MetroCluster esistono tra i siti per i quali le comunicazioni nodo-nodo e l'accesso al disco si basano sulla connettività di rete site-to-site. La capacità di monitorare il battito cardiaco del resto del cluster è limitata. AUSO deve discriminare tra una situazione in cui l'altro sito è effettivamente inattivo piuttosto che non disponibile a causa di un problema di rete.

Di conseguenza, un controller in una coppia ha può richiedere un takeover se rileva un guasto del controller verificatosi per un motivo specifico, ad esempio un panico del sistema. Può anche richiedere un takeover in caso di perdita totale della connettività, talvolta nota come battito cardiaco perso.

Un sistema MetroCluster può eseguire uno switchover automatico in modo sicuro solo quando viene rilevato un guasto specifico nel sito originale. Inoltre, il controller che prende la proprietà del sistema di storage deve essere in grado di garantire che i dati su disco e NVRAM siano sincronizzati. Il controller non è in grado di garantire la sicurezza di uno switchover solo perché ha perso il contatto con il sito di origine, cosa che potrebbe essere ancora operativa. Per ulteriori opzioni per automatizzare uno switchover, vedere le informazioni sulla soluzione MetroCluster Tiebreaker (MCTB) nella sezione successiva.

Tiebreaker MetroCluster con MetroCluster fabric-attached

Il "Tiebreaker NetApp MetroCluster" software può essere eseguito su un terzo sito per monitorare lo stato dell'ambiente MetroCluster, inviare notifiche e, facoltativamente, imporre uno switchover in una situazione di emergenza. Una descrizione completa di Tiebreaker "Sito di supporto NetApp"è disponibile sul , ma lo scopo principale di MetroCluster Tiebreaker è quello di rilevare la perdita del sito. Inoltre, deve discriminare tra la perdita del sito e la perdita della connettività. Ad esempio, lo switchover non deve essere eseguito perché il tiebreaker non è riuscito a raggiungere il sito primario; questo spiega perché il tiebreaker monitora anche la capacità del sito remoto di contattare il sito primario.

Lo switchover automatico con AUSO è compatibile anche con l'MCTB. AUSO reagisce in modo molto rapido perché è progettato per rilevare eventi di errore specifici e quindi richiamare lo switchover solo quando i plex NVRAM e SyncMirror sono sincronizzati.

Al contrario, il Tiebreaker è localizzato a distanza e quindi deve attendere che un temporizzatore trascorra prima di dichiarare un sito morto. Il tiebreaker alla fine rileva il tipo di guasto del controller coperto da AUSO, ma in generale AUSO ha già avviato lo switchover e, eventualmente, ha completato lo switchover prima che il tiebreaker agisca. Il secondo comando switchover risultante proveniente dal tiebreaker verrebbe rifiutato.

Il software MCTB non verifica se NVRAM era e/o i plex sono sincronizzati quando si forza uno switchover. Lo switchover automatico, se configurato, deve essere disattivato durante le attività di manutenzione che causano una perdita di sincronizzazione dei plessi NVRAM o SyncMirror.

Inoltre, l'MCTB potrebbe non risolvere un disastro continuo che porta alla seguente sequenza di eventi:

1. La connettività tra i siti viene interrotta per più di 30 secondi.
2. Timeout della replica SyncMirror e proseguimento delle operazioni sul sito primario, lasciando inattiva la replica remota.
3. Il sito primario viene perso. Il risultato è la presenza di modifiche non replicate sul sito primario. Uno switchover potrebbe quindi essere indesiderato per una serie di motivi, tra cui:
    - I dati critici potrebbero essere presenti sul sito primario e quindi ripristinabili. Uno switchover che ha permesso all'applicazione di continuare a funzionare eliminava efficacemente i dati critici.
    - Un'applicazione sul sito rimasto che stava utilizzando le risorse di storage sul sito primario al momento della perdita del sito potrebbe avere memorizzato nella cache i dati. Uno switchover introdurrebbe una versione obsoleta dei dati che non corrisponde alla cache.
    - Un sistema operativo del sito rimasto che utilizzava le risorse di storage del sito primario al momento della perdita del sito potrebbe avere memorizzato i dati nella cache. Uno switchover introdurrebbe una versione obsoleta dei dati che non corrisponde alla cache. L'opzione più sicura è configurare tiebreaker in modo da inviare un avviso se rileva un guasto del sito e chiedere a una persona di decidere se forzare uno switchover. Potrebbe essere necessario arrestare le applicazioni e/o i sistemi operativi per cancellare i dati memorizzati nella cache. Inoltre, è possibile utilizzare le impostazioni NVFAIL per aggiungere ulteriore protezione e semplificare il processo di failover.

ONTAP Mediator con MetroCluster IP

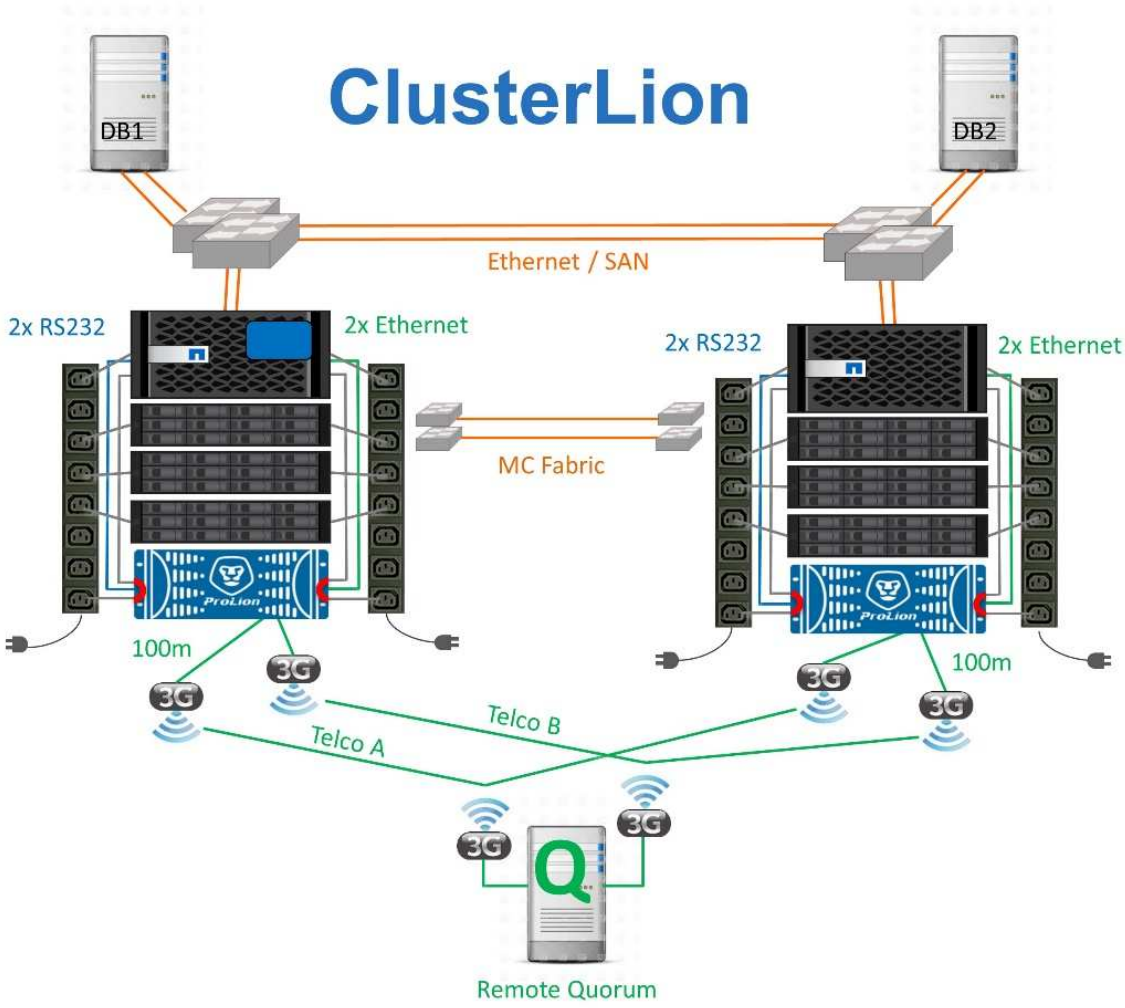
ONTAP Mediator viene utilizzato con MetroCluster IP e con alcune altre soluzioni ONTAP. Funziona come un servizio di tiebreaker tradizionale, proprio come il software MetroCluster Tiebreaker descritto in precedenza, ma include anche una funzione critica: Eseguire uno switchover automatizzato e non assistito.

Un MetroCluster fabric-attached ha accesso diretto ai dispositivi di storage del sito opposto. Ciò consente a un controller MetroCluster di monitorare lo stato degli altri controller leggendo i dati heartbeat dalle unità. In questo modo, un controller riconosce il guasto di un altro controller ed esegue uno switchover.

Al contrario, l'architettura IP di MetroCluster instrada tutti i/o esclusivamente attraverso la connessione controller-controller; non vi è accesso diretto ai dispositivi di storage sul sito remoto. Questo limita la possibilità per un controller di rilevare gli errori ed eseguire uno switchover. Pertanto, come dispositivo di tiebreaker occorre il ONTAP Mediator per rilevare la perdita di un sito ed eseguire automaticamente uno switchover.

Terzo sito virtuale con ClusterLion

ClusterLion è un'appliance di monitoraggio MetroCluster avanzata che funziona come un terzo sito virtuale. Questo approccio consente di implementare MetroCluster in maniera sicura in una configurazione a due siti con una funzionalità di switchover completamente automatizzata. Inoltre, ClusterLion può eseguire ulteriori operazioni di monitoraggio a livello di rete ed eseguire operazioni post-switchover. La documentazione completa è disponibile presso ProLion.



- Gli appliance ClusterLion monitorano lo stato dei controller con cavi Ethernet e seriali collegati direttamente.
- I due dispositivi sono collegati tra loro mediante connessioni wireless 3G ridondanti.
- L'alimentazione alla centralina ONTAP viene instradata attraverso i relè interni. In caso di guasto a un sito, ClusterLion, che contiene un sistema UPS interno, interrompe i collegamenti di alimentazione prima di richiamare uno switchover. Questo processo assicura che non si verifichi alcuna condizione split-brain.
- ClusterLion esegue uno switchover entro il timeout SyncMirror di 30 secondi o non lo esegue affatto.
- ClusterLion non esegue uno switchover a meno che gli stati della NVRAM e dei plex SyncMirror non siano sincronizzati.
- Poiché ClusterLion esegue uno switchover solo se MetroCluster è completamente sincronizzato, NVFAIL non è necessario. Questa configurazione consente ad ambienti che si estendono tra diversi siti, come un Oracle RAC esteso, di rimanere online anche durante uno switchover non pianificato.
- Il supporto include MetroCluster fabric-attached e MetroCluster IP

SyncMirror

La base della protezione dei dati di Oracle con un sistema MetroCluster è SyncMirror, una tecnologia di mirroring sincrono scale-out dalle performance massime.

Data Protection con SyncMirror

Al livello più semplice, la replica sincrona significa che qualsiasi modifica deve essere apportata a entrambi i lati dello storage con mirroring prima che venga riconosciuta. Ad esempio, se un database sta scrivendo un registro o un guest VMware viene aggiornato, non deve mai andare persa una scrittura. Come livello di protocollo, il sistema di storage non deve riconoscere la scrittura fino a quando non è stato assegnato a un supporto non volatile in entrambi i siti. Solo allora è sicuro procedere senza il rischio di perdita dei dati.

L'utilizzo di una tecnologia di replica sincrona è il primo passo nella progettazione e nella gestione di una soluzione di replica sincrona. La considerazione più importante è capire cosa potrebbe accadere durante i vari scenari di guasto pianificati e non pianificati. Non tutte le soluzioni di replica sincrona offrono le stesse funzionalità. Se hai bisogno di una soluzione che offra un recovery point objective (RPO) pari a zero, ovvero zero data loss, devi prendere in considerazione tutti gli scenari di guasto. In particolare, qual è il risultato previsto quando la replica è impossibile a causa della perdita di connettività tra i siti?

Disponibilità dei dati SyncMirror

La replica MetroCluster si basa sulla tecnologia NetApp SyncMirror, che è progettata per passare in modo efficiente dalla modalità sincrona alla modalità sincrona e viceversa. Questa funzionalità soddisfa i requisiti dei clienti che richiedono una replica sincrona, ma che hanno bisogno anche di un'alta disponibilità per i propri servizi dati. Ad esempio, se la connettività a un sito remoto viene interrotta, è generalmente preferibile che il sistema di archiviazione continui a funzionare in uno stato non replicato.

Molte soluzioni di replica sincrona sono in grado di funzionare solo in modalità sincrona. Questo tipo di replica "tutto o niente" viene talvolta chiamato modalità domino. Tali sistemi storage smettono di fornire i dati piuttosto che permettere che le copie locali e remote dei dati diventino non sincronizzate. Se la replica viene forzata, la risincronizzazione può richiedere molto tempo e lasciare un cliente esposto a una perdita di dati completa durante il tempo in cui il mirroring viene ristabilita.

Non solo SyncMirror può passare alla modalità sincrona senza problemi se il sito remoto non è raggiungibile, ma può anche risincronizzare rapidamente uno stato RPO = 0 al ripristino della connettività. La copia obsoleta dei dati nel sito remoto può anche essere preservata in uno stato utilizzabile durante la risincronizzazione, garantendo l'esistenza in ogni momento di copie locali e remote dei dati.

Quando è richiesta la modalità domino, NetApp offre SnapMirror Synchronous (SM-S). Esistono anche opzioni a livello di applicazione, come Oracle DataGuard o SQL Server Always on Availability Groups. Il mirroring del disco a livello del sistema operativo può essere opzionale. Per ulteriori informazioni e opzioni, consulta il tuo NetApp o il partner account team.

MetroCluster e NVFAIL

NVFAIL è una funzionalità generale di integrità dei dati di ONTAP progettata per massimizzare la protezione dell'integrità dei dati con i database.

Questa sezione espande la spiegazione di ONTAP NVFAIL di base per affrontare argomenti specifici di MetroCluster.

Con MetroCluster, una scrittura non viene riconosciuta fino a quando non è stata registrata nella NVRAM locale e nella NVRAM su almeno un altro controller. Questo approccio garantisce che un guasto dell'hardware o un'interruzione di corrente non comporti la perdita dell'i/o in-flight Se si verifica un guasto nella NVRAM locale o nella connettività ad altri nodi, i dati non verranno più mirrorati.

Se la NVRAM locale riporta un errore, il nodo si arresta. Questo arresto determina il failover su un partner controller quando vengono utilizzate coppie ha. Con MetroCluster, il comportamento dipende dalla

configurazione complessiva scelta, ma può portare al failover automatico della nota remota. In ogni caso, nessun dato viene perso perché il controller che subisce l'errore non ha confermato l'operazione di scrittura.

Un guasto di connettività site-to-site che blocca la replica NVRAM ai nodi remoti è una situazione più complicata. Le scritture non vengono più replicate sui nodi remoti, con la possibilità di perdita di dati in caso di errore catastrofico su un controller. Cosa più importante, il tentativo di failover su un nodo diverso in queste condizioni comporta una perdita di dati.

Il fattore di controllo è se la NVRAM è sincronizzata. Se la NVRAM è sincronizzata, il failover da nodo a nodo può procedere in tutta sicurezza senza il rischio di perdita di dati. In una configurazione MetroCluster, se la NVRAM e i plessi degli aggregati sottostanti sono sincronizzati, è possibile effettuare lo switchover senza correre il rischio di perdita di dati.

ONTAP non consente alcun failover o switchover quando i dati non sono sincronizzati, a meno che non sia forzato il failover o lo switchover. La forzatura di una modifica delle condizioni in questo modo riconosce che i dati potrebbero essere lasciati indietro nel controllore originale e che la perdita di dati è accettabile.

I database sono particolarmente vulnerabili al danneggiamento se un failover o uno switchover è forzato, perché mantengono cache interne di dati su disco di dimensioni maggiori. In caso di failover o switchover forzato, le modifiche riconosciute in precedenza vengono eliminate del tutto. Il contenuto dell'array di storage torna indietro nel tempo e lo stato della cache del database non riflette più lo stato dei dati su disco.

Per proteggere le applicazioni da questa situazione, ONTAP consente di configurare i volumi per una protezione speciale contro gli errori della NVRAM. Quando attivato, questo meccanismo di protezione determina l'ingresso di un volume nello stato chiamato NVFAIL. Questo stato causa errori di i/o che causano l'arresto di un'applicazione in modo che non utilizzino dati obsoleti. I dati non devono essere persi perché eventuali scritture riconosciute sono ancora presenti nel sistema di storage e, nel caso dei database, tutti i dati delle transazioni con commit devono essere presenti nei registri.

Solitamente, gli amministratori dovranno arrestare completamente gli host prima di riportare manualmente LUN e volumi in linea. Sebbene queste fasi possano comportare un certo lavoro, questo approccio è il modo più sicuro per garantire l'integrità dei dati. Non tutti i dati richiedono questa protezione, motivo per cui il comportamento di NVFAIL può essere configurato in base al volume.

NVFAIL forzato manualmente

L'opzione più sicura per forzare uno switchover con un cluster di applicazioni (inclusi VMware, Oracle RAC e altri) distribuito tra i siti dipende da come specificato `-force-nvfail-all` alla riga di comando. Questa opzione è disponibile come misura di emergenza per assicurarsi che tutti i dati memorizzati nella cache vengano eliminati. Se un host utilizza risorse di storage situate originariamente nel sito colpito da disastro, riceve errori di i/o o un handle di file obsoleto (`ESTALE`). I database Oracle si arrestano in modo anomalo e i file system possono andare completamente offline o passare alla modalità di sola lettura.

Al termine dello switchover, il `in-nvfailed-state` Il flag deve essere cancellato e i LUN devono essere messi online. Al termine di questa attività, è possibile riavviare il database. È possibile automatizzare queste attività per ridurre l'RTO.

dr-force-nvfail

Come misura di sicurezza generale, impostare `dr-force-nvfail` contrassegnare tutti i volumi a cui è possibile accedere da un sito remoto durante le normali operazioni, ovvero si tratta di attività utilizzate prima del failover. Il risultato di questa impostazione è che i volumi remoti selezionati diventano non disponibili quando vengono immessi `in-nvfailed-state` durante uno switchover. Al termine dello switchover, il `in-nvfailed-state` Il flag deve essere cancellato e i LUN devono essere messi online. Al termine di queste attività, è possibile riavviare le applicazioni. È possibile automatizzare queste attività per ridurre l'RTO.

Il risultato è come usare l' `-force-nvfail-all` flag per commutatori manuali. Tuttavia, il numero di volumi interessati può essere limitato solo a quei volumi che devono essere protetti da applicazioni o sistemi operativi con cache obsolete.

Ci sono due requisiti critici per un ambiente che non utilizza `dr-force-nvfail` su volumi applicativi:

- Uno switchover forzato non deve avvenire più di 30 secondi dopo la perdita del sito primario.
- Lo switchover non deve essere eseguito durante le attività di manutenzione o in altre condizioni in cui i plex SyncMirror o la replica della NVRAM non sono sincronizzati. Il primo requisito può essere soddisfatto con il software tiebreaker configurato per eseguire uno switchover entro 30 secondi da un guasto del sito. Questo requisito non significa che lo switchover debba essere eseguito entro 30 secondi dal rilevamento di un guasto del sito. Ciò significa che non è più sicuro forzare uno switchover se sono trascorsi 30 secondi da quando un sito è stato confermato operativo.

Il secondo requisito può essere parzialmente soddisfatto disattivando tutte le funzionalità di switchover automatico quando la configurazione di MetroCluster non è sincronizzata. Un'opzione migliore è quella di disporre di una soluzione di tiebreaker in grado di monitorare lo stato di salute della replica NVRAM e dei plessi SyncMirror. Se il cluster non è completamente sincronizzato, il tiebreaker non deve attivare uno switchover.

Il software NetApp MCTB non è in grado di monitorare lo stato di sincronizzazione, pertanto deve essere disattivato quando MetroCluster non è sincronizzato per alcun motivo. ClusterLion include funzionalità di monitoraggio NVRAM e plex e può essere configurato in modo da non attivare lo switchover a meno che il sistema MetroCluster non sia confermato completamente sincronizzato.

Istanza singola di Oracle

Come indicato in precedenza, la presenza di un sistema MetroCluster non implica necessariamente l'aggiunta o la modifica delle Best practice per l'utilizzo di un database. La maggior parte dei database attualmente in esecuzione sui sistemi MetroCluster dei clienti è a singola istanza e segue le raccomandazioni contenute nella documentazione relativa a Oracle su ONTAP.

Failover con un sistema operativo preconfigurato

SyncMirror fornisce una copia sincrona dei dati nel sito di disaster recovery, ma per renderli disponibili sono necessari un sistema operativo e le applicazioni associate. L'automazione di base può migliorare notevolmente il tempo di failover dell'ambiente complessivo. I prodotti Clusterware come Veritas Cluster Server (VCS) vengono spesso utilizzati per creare un cluster in tutti i siti e in molti casi il processo di failover può essere guidato con semplici script.

In caso di perdita dei nodi primari, il clusterware (o gli script) viene configurato in modo da portare i database online nel sito alternativo. Un'opzione è creare server di standby preconfigurati per le risorse NFS o SAN che compongono il database. Se il sito primario non funziona, il clusterware o l'alternativa con script esegue una sequenza di azioni simile alle seguenti:

1. Forzare uno switchover su MetroCluster
2. Rilevamento di LUN FC (solo SAN)
3. Montaggio di file system e/o montaggio di gruppi di dischi ASM
4. Avvio del database

Il requisito principale di questo approccio è rappresentato da un sistema operativo in esecuzione sul sito remoto. Deve essere preconfigurato con i file binari di Oracle, il che significa anche che attività come l'applicazione delle patch Oracle devono essere eseguite sul sito primario e di standby. In alternativa, è possibile eseguire il mirroring dei file binari di Oracle nel sito remoto e montarli se viene dichiarato un disastro.

La procedura di attivazione effettiva è semplice. Comandi come il rilevamento delle LUN richiedono solo pochi comandi per ogni porta FC. Il montaggio del file system non è altro che un `mount` E sia i database che ASM possono essere avviati e arrestati dalla CLI con un unico comando. Se i volumi e i file system non vengono utilizzati nel sito di disaster recovery prima dello switchover, non è necessario impostare alcun requisito `dr-force- nvfail` sui volumi.

Failover con un sistema operativo virtualizzato

Il failover degli ambienti di database può essere esteso per includere il sistema operativo stesso. In teoria, questo failover può essere eseguito con le LUN di avvio, ma nella maggior parte dei casi con un sistema operativo virtualizzato. La procedura è simile ai seguenti passaggi:

1. Forzare uno switchover su MetroCluster
2. Montaggio dei datastore che ospitano le macchine virtuali del server di database
3. Avvio delle macchine virtuali
4. Avvio manuale dei database o configurazione delle macchine virtuali per avviare automaticamente i database, ad esempio, un cluster ESX può estendersi su diversi siti. In caso di disastro, dopo lo switchover, è possibile portare online le macchine virtuali nel sito di disaster recovery. Fino a quando i datastore che ospitano i database server virtualizzati non saranno in uso in occasione di un evento di emergenza, non sarà necessario impostare alcun valore `dr-force- nvfail` sui volumi associati.

Oracle Extended RAC

Molti clienti ottimizzano il proprio RTO estendendo un cluster Oracle RAC tra i vari siti, ottenendo una configurazione completamente Active-Active. La progettazione complessiva diventa più complicata perché deve includere la gestione del quorum di Oracle RAC. Inoltre, entrambi i siti accedono ai dati, il che significa che uno switchover forzato può portare all'utilizzo di una copia dei dati non aggiornata.

Sebbene una copia dei dati sia presente in entrambi i siti, solo il controller attualmente proprietario di un aggregato può fornire i dati. Pertanto, con i cluster RAC estesi, i nodi remoti devono eseguire l'i/o attraverso una connessione site-to-site. Il risultato è un'aggiunta di latenza i/o, ma generalmente questa latenza non rappresenta un problema. Anche la rete di interconnessione RAC deve essere estesa su più siti, il che significa che è comunque necessaria una rete ad alta velocità e a bassa latenza. Se la latenza aggiunta causa un problema, il cluster può essere azionato in maniera Active-passive. Quindi, le operazioni i/o-intensive devono essere indirizzate ai nodi RAC locali del controller proprietario degli aggregati. I nodi remoti eseguono quindi operazioni i/o più chiare o vengono utilizzati esclusivamente come server warm standby.

Se è richiesto un RAC esteso Active-Active, la sincronizzazione attiva di SnapMirror deve essere presa in considerazione al posto di MetroCluster. La replica SM-As consente di preferire una replica specifica dei dati. Pertanto, può essere integrato un cluster RAC esteso in cui tutte le letture avvengono localmente. Gli i/o in lettura non attraversano mai i siti, offrendo la minore latenza possibile. Tutte le attività di scrittura devono comunque transito sulla connessione tra siti, ma tale traffico è inevitabile con qualsiasi soluzione di mirroring sincrono.

Se con Oracle RAC vengono utilizzate LUN di avvio, compresi i dischi di avvio virtualizzati, `misscount` potrebbe essere necessario modificare il parametro. Per ulteriori informazioni sui parametri di timeout RAC, vedere "Oracle RAC con ONTAP".

Configurazione a due siti

Una configurazione RAC estesa a due siti può fornire servizi di database Active-Active che possono sopravvivere a molti scenari ma non a tutti.

File di voto RAC

La prima considerazione da prendere in considerazione per la distribuzione di RAC esteso su MetroCluster deve essere la gestione del quorum. Oracle RAC dispone di due meccanismi per gestire il quorum: Heartbeat del disco e heartbeat della rete. L'heartbeat del disco controlla l'accesso allo storage utilizzando i file di voto. Con una configurazione RAC a sito singolo, una singola risorsa di voto è sufficiente fintanto che il sistema storage sottostante offre funzionalità ha.

Nelle versioni precedenti di Oracle, i file di voto erano posizionati su dispositivi di archiviazione fisici, ma nelle versioni correnti di Oracle i file di voto sono memorizzati in gruppi di dischi ASM.

Oracle RAC è supportato con NFS. Durante il processo di installazione della griglia, viene creata una serie di processi ASM per presentare la posizione NFS utilizzata per i file della griglia come un gruppo di dischi ASM. Il processo è quasi trasparente per l'utente finale e non richiede alcuna gestione ASM continua al termine dell'installazione.

Il primo requisito di una configurazione a due siti è garantire che ogni sito possa sempre accedere a più della metà dei file di voto in modo da garantire un processo di disaster recovery senza interruzioni. Questa attività era semplice prima che i file di voto fossero memorizzati in gruppi di dischi ASM, ma oggi gli amministratori devono comprendere i principi di base della ridondanza ASM.

I gruppi di dischi ASM hanno tre opzioni di ridondanza `external`, `normal`, e. `high`. In altre parole, senza mirror, con mirroring e a 3 vie con mirroring. Un'opzione più recente chiamata `Flex` è anche disponibile, ma raramente utilizzato. Il livello di ridondanza e il posizionamento dei dispositivi ridondanti controllano ciò che accade negli scenari di errore. Ad esempio:

- Posizionamento dei file di votazione su un `diskgroup` con `external` la risorsa di ridondanza garantisce l'eliminazione di un sito se la connettività tra siti viene persa.
- Posizionamento dei file di votazione su un `diskgroup` con `normal` La ridondanza con un solo disco ASM per sito garantisce l'eliminazione dei nodi su entrambi i siti se la connettività tra i siti viene persa perché nessuno dei due siti dispone di un quorum di maggioranza.
- Posizionamento dei file di votazione su un `diskgroup` con `high` la ridondanza con due dischi su un sito e un singolo disco sull'altro sito consente operazioni active-active quando entrambi i siti sono operativi e reciprocamente raggiungibili. Tuttavia, se il sito a disco singolo è isolato dalla rete, il sito viene eliminato.

Heartbeat rete RAC

L'heartbeat della rete Oracle RAC monitora la raggiungibilità dei nodi in tutta l'interconnessione cluster. Per rimanere nel cluster, un nodo deve essere in grado di contattare più della metà degli altri nodi. In un'architettura a due siti, questo requisito crea le seguenti scelte per il numero di nodi RAC:

- Il posizionamento di un numero uguale di nodi per sito comporta l'espulsione in un sito nel caso in cui la connettività di rete venga persa.

- Il posizionamento di N nodi su un sito e N+1 nodi sul sito opposto garantisce che la perdita di connettività intersito determini nel sito con il maggior numero di nodi rimanenti nel quorum di rete e nel sito con meno nodi evicting.

Prima di Oracle 12cR2, non era fattibile controllare quale lato avrebbe subito un'eviction durante la perdita del sito. Quando ogni sito ha un numero uguale di nodi, l'evocazione è controllata dal nodo master, che in generale è il primo nodo RAC da avviare.

Oracle 12cR2 introduce la funzionalità di ponderazione dei nodi. Questa funzionalità consente agli amministratori di controllare in che modo Oracle risolve le condizioni split-brain. Ad esempio, il seguente comando imposta la preferenza per un nodo specifico in un RAC:

```
[root@host-a ~]# /grid/bin/crsctl set server css_critical yes
CRS-4416: Server attribute 'CSS_CRITICAL' successfully changed. Restart
Oracle High Availability Services for new value to take effect.
```

Dopo aver riavviato Oracle High-Availability Services, la configurazione si presenta come segue:

```
[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
```

Nodo `host-a` è ora designato come server critico. Se i due nodi RAC sono isolati, `host-a` sopravvive, e. `host-b` è sfrattato.

Per informazioni dettagliate, consultare il white paper Oracle "Panoramica tecnica su Oracle Clusterware 12c Release 2. "

Per le versioni di Oracle RAC precedenti a 12cR2, il nodo master può essere identificato controllando i registri CRS come segue:

```
[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
```

Questo registro indica che il nodo master è `2` e il nodo `host-a` Ha un ID di `1`. Questo significa che `host-a` non è il nodo master. L'identità del nodo master può essere confermata con il comando `olsnodes -n`.

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

Il nodo con un ID di `2` è `host-b`, che è il nodo master. In una configurazione con un numero uguale di nodi su ogni sito, il sito con `host-b` è il sito che sopravvive se i due set perdono la connettività di rete per qualsiasi motivo.

È possibile che la voce di log che identifica il nodo master rimanga fuori dal sistema. In questa situazione, è possibile utilizzare i timestamp dei backup OCR (Oracle Cluster Registry).

```
[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
```

Questo esempio mostra che il nodo master è `host-b`. Indica anche una modifica nel nodo master da `host-a` a. `host-b` Da qualche parte tra il 2:05 e il 21:39 maggio 4. Questo metodo di identificazione del nodo master è sicuro da utilizzare solo se sono stati controllati anche i log CRS, poiché è possibile che il nodo master sia

cambiato dal precedente backup OCR. Se questa modifica si è verificata, dovrebbe essere visibile nei registri OCR.

La maggior parte dei clienti sceglie un singolo gruppo di dischi di voto che gestisce l'intero ambiente e un numero uguale di nodi RAC su ciascun sito. Il gruppo di dischi deve essere collocato nel sito che contiene il database. Il risultato è che la perdita di connettività provoca sfratto sul sito remoto. Il sito remoto non dispone più del quorum né avrebbe accesso ai file di database, ma il sito locale continua a funzionare normalmente. Quando la connettività viene ripristinata, l'istanza remota può essere riportata nuovamente in linea.

In caso di emergenza, è necessario uno switchover per portare online i file di database e il gruppo di dischi di voto sul sito rimasto. Se il disastro consente AD AUSO di attivare lo switchover, NVFAIL non viene attivato perché il cluster è sincronizzato e le risorse di storage vengono normalmente online. AUSO è un'operazione molto veloce e dovrebbe essere completata prima del `disktimeout` il periodo scade.

Poiché ci sono solo due siti, non è possibile utilizzare alcun tipo di software di rottura automatica esterna, il che significa che lo switchover forzato deve essere un'operazione manuale.

Configurazioni a tre siti

Un cluster RAC esteso è molto più semplice da progettare con tre siti. I due siti che ospitano ciascuna metà del sistema MetroCluster supportano anche i carichi di lavoro del database, mentre il terzo sito funge da tiebreaker sia per il database che per il sistema MetroCluster. La configurazione di Oracle Tiebreaker può essere semplice come collocare un membro del gruppo di dischi ASM utilizzato per il voto su un sito 3rd e può anche includere un'istanza operativa sul sito 3rd per garantire che vi sia un numero dispari di nodi nel cluster RAC.

> Per informazioni importanti sull'utilizzo di NFS in una configurazione RAC estesa, consultare la documentazione Oracle relativa al "gruppo di errori del quorum". In sintesi, potrebbe essere necessario modificare le opzioni di montaggio NFS per includere l'opzione soft per garantire che la perdita di connettività alle risorse quorum di hosting del sito 3rd non blocchi i server Oracle primari o i processi Oracle RAC.

Sincronizzazione attiva di SnapMirror

Panoramica

SnapMirror Active Sync ti consente di creare ambienti di database Oracle a disponibilità ultra elevata in cui le LUN sono disponibili da due cluster di storage diversi.

Con la sincronizzazione attiva di SnapMirror, non vi è alcuna copia "primaria" e "secondaria" dei dati. Ogni cluster può fornire i/o in lettura dalla propria copia locale dei dati e ciascun cluster replicherà una scrittura al partner. Il risultato è un comportamento io simmetrico.

Tra le altre opzioni, questo consente di eseguire Oracle RAC come cluster esteso con istanze operative su entrambi i siti. In alternativa, è possibile creare cluster di database attivi-passivi RPO=0 in cui è possibile spostare i database a singola istanza tra i siti durante un'interruzione del sito e questo processo può essere automatizzato tramite prodotti come Pacemaker o VMware ha. La base di queste opzioni è la replica sincrona gestita dalla sincronizzazione attiva di SnapMirror.

Replica sincrona

Durante il funzionamento normale, la sincronizzazione attiva di SnapMirror fornisce sempre RPO=0 replica sincrona, con un'eccezione. Se i dati non possono essere replicati, ONTAP rilascerà il requisito di replicare i dati e riprendere la distribuzione io su un sito mentre le LUN dell'altro sito vengono portate offline.

Hardware per lo storage

A differenza di altre soluzioni di disaster recovery per lo storage, SnapMirror Active Sync offre una flessibilità asimmetrica della piattaforma. Non è necessario che l'hardware di ciascun sito sia identico. Questa funzionalità consente di dimensionare correttamente l'hardware utilizzato per supportare la sincronizzazione attiva di SnapMirror. Il sistema di storage remoto può essere identico al sito primario se deve supportare un carico di lavoro di produzione completo, ma se un disastro determina una riduzione dell'i/o, rispetto a un sistema più piccolo nel sito remoto potrebbe risultare più conveniente.

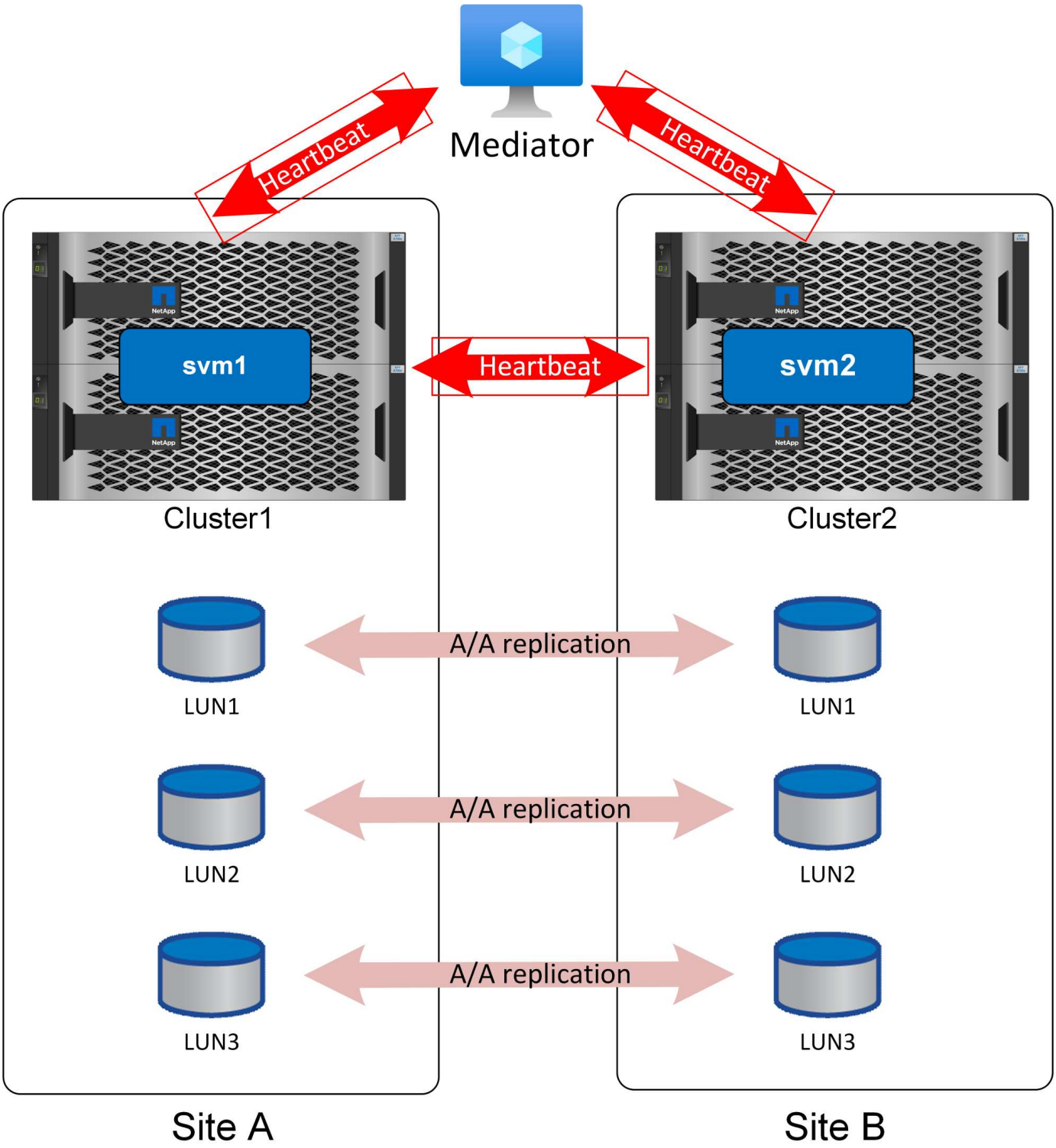
Mediatore ONTAP

ONTAP Mediator è un'applicazione software che viene scaricata dal supporto NetApp e che viene in genere distribuita su una piccola macchina virtuale. Il ONTAP Mediator non è un Tiebreaker quando viene utilizzato con la sincronizzazione attiva di SnapMirror. È un canale di comunicazione alternativo per i due cluster che partecipano alla replica sincrona attiva SnapMirror. Le operazioni automatizzate sono gestite da ONTAP in base alle risposte ricevute dal partner tramite connessioni dirette e tramite il mediatore.

Mediatore ONTAP

Il mediatore è necessario per automatizzare in modo sicuro il failover. Idealmente, sarebbe posizionato su un sito 3rd indipendente, ma può comunque funzionare per la maggior parte delle esigenze se colocated con uno dei cluster che partecipano alla replica.

Il mediatore non è propriamente uno strumento decisivo, sebbene questa sia effettivamente la sua funzione. Il mediatore aiuta a determinare lo stato dei nodi del cluster e supporta il processo di commutazione automatica in caso di guasto del sito. Il mediatore non trasferisce dati in nessun caso.



La sfida #1 con il failover automatizzato è il problema split-brain, e questo problema sorge se i due siti perdono la connettività tra loro. Che cosa dovrebbe accadere? Non si desidera che due siti diversi si designino come copie sopravvissute dei dati, ma in che modo un singolo sito può distinguere tra la perdita effettiva del sito opposto e l'impossibilità di comunicare con il sito opposto?

Qui entra il mediatore nell'immagine. Se si trova in un sito 3rd e ciascun sito dispone di una connessione di rete separata per tale sito, è disponibile un percorso aggiuntivo per ciascun sito per convalidare lo stato dell'altro. Esaminare nuovamente l'immagine sopra riportata e considerare i seguenti scenari.

- Cosa succede se il mediatore non riesce o è irraggiungibile da uno o entrambi i siti?
  - I due cluster possono ancora comunicare tra loro sullo stesso link utilizzato per i servizi di replica.
  - I dati sono ancora serviti con protezione RPO=0/7
- Cosa succede se il sito A non funziona?
  - Il sito B vedrà che entrambi i canali di comunicazione si interrompono.
  - Il sito B sostituirà i servizi dati, ma senza RPO = mirroring 0:1
- Cosa succede se il sito B non funziona?
  - Il sito A vedrà che entrambi i canali di comunicazione si interrompono.
  - Il sito A sostituirà i servizi dati, ma senza RPO = mirroring 0:1

Esiste un altro scenario da considerare: La perdita del collegamento di replica dei dati. In caso di perdita del link di replica tra i siti, RPO=0 Mirroring sarà ovviamente impossibile. Che cosa dovrebbe accadere allora?

Questo è controllato dallo stato del sito preferito. In una relazione SM-AS, uno dei siti è secondario all'altro. Questo non ha alcun effetto sulle normali operazioni e tutto l'accesso ai dati è simmetrico, ma se la replica viene interrotta, il legame dovrà essere interrotto per riprendere le operazioni. Ne risulta che il sito preferito continuerà le operazioni senza mirroring e il sito secondario interromperà l'elaborazione io fino al ripristino della comunicazione di replica.

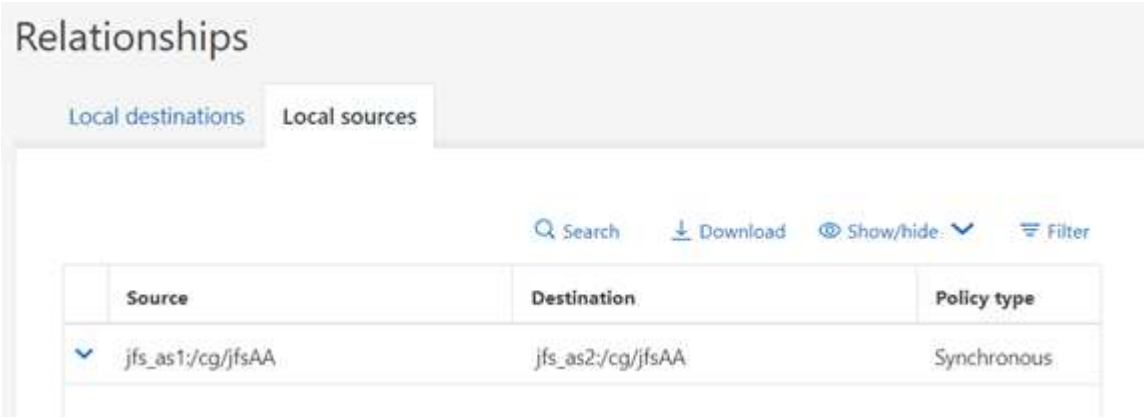
Sito preferito sincronizzazione attiva SnapMirror

Il comportamento di sincronizzazione attiva di SnapMirror è simmetrico, con una eccezione importante: Configurazione del sito preferita.

La sincronizzazione attiva di SnapMirror considererà un sito la "fonte" e l'altro la "destinazione". Ciò implica una relazione di replica unidirezionale, ma ciò non si applica al comportamento io. La replica è bidirezionale e simmetrica, mentre i tempi di risposta io sono identici su entrambi i lati del mirror.

La `source` designazione è controlla il sito preferito. In caso di perdita del link di replica, i percorsi delle LUN nella copia di origine continueranno a fornire i dati mentre i percorsi delle LUN nella copia di destinazione non saranno disponibili finché la replica non viene ristabilita e SnapMirror ritorna allo stato sincrono. I percorsi riprenderanno a fornire i dati.

La configurazione di origine/destinazione può essere visualizzata tramite SystemManager:



O all'interfaccia 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
```

La chiave è che source è l'SVM su cluster1. Come menzionato sopra, i termini "origine" e "destinazione" non descrivono il flusso di dati replicati. Entrambi i siti possono elaborare una scrittura e replicarla nel sito opposto. In effetti, entrambi i cluster sono origini e destinazioni. L'effetto della designazione di un cluster come origine controlla semplicemente quale cluster sopravvive come sistema di storage in lettura e scrittura in caso di perdita del link di replica.

Topologia di rete

Accesso uniforme

Una rete ad accesso uniforme significa che gli host sono in grado di accedere ai percorsi su entrambi i siti (o ai domini di errore all'interno dello stesso sito).

Una caratteristica importante di SM-AS è la capacità di configurare i sistemi storage per sapere dove si trovano gli host. Quando si mappano i LUN a un determinato host, è possibile indicare se sono prossimali o meno a un determinato sistema di archiviazione.

Impostazioni di prossimità

La prossimità si riferisce a una configurazione per cluster che indica che un determinato WWN host o ID iniziatore iSCSI appartiene a un host locale. Si tratta di un secondo passo opzionale per la configurazione dell'accesso LUN.

Il primo passo è la normale configurazione di igroup. Ogni LUN deve essere mappato a un igroup che contiene gli ID WWN/iSCSI degli host che devono accedere a quel LUN. Controlla quale host ha *accesso* a un LUN.

Il secondo passo, opzionale, consiste nel configurare la prossimità all'host. Questo non controlla l'accesso, controlla *priority*.

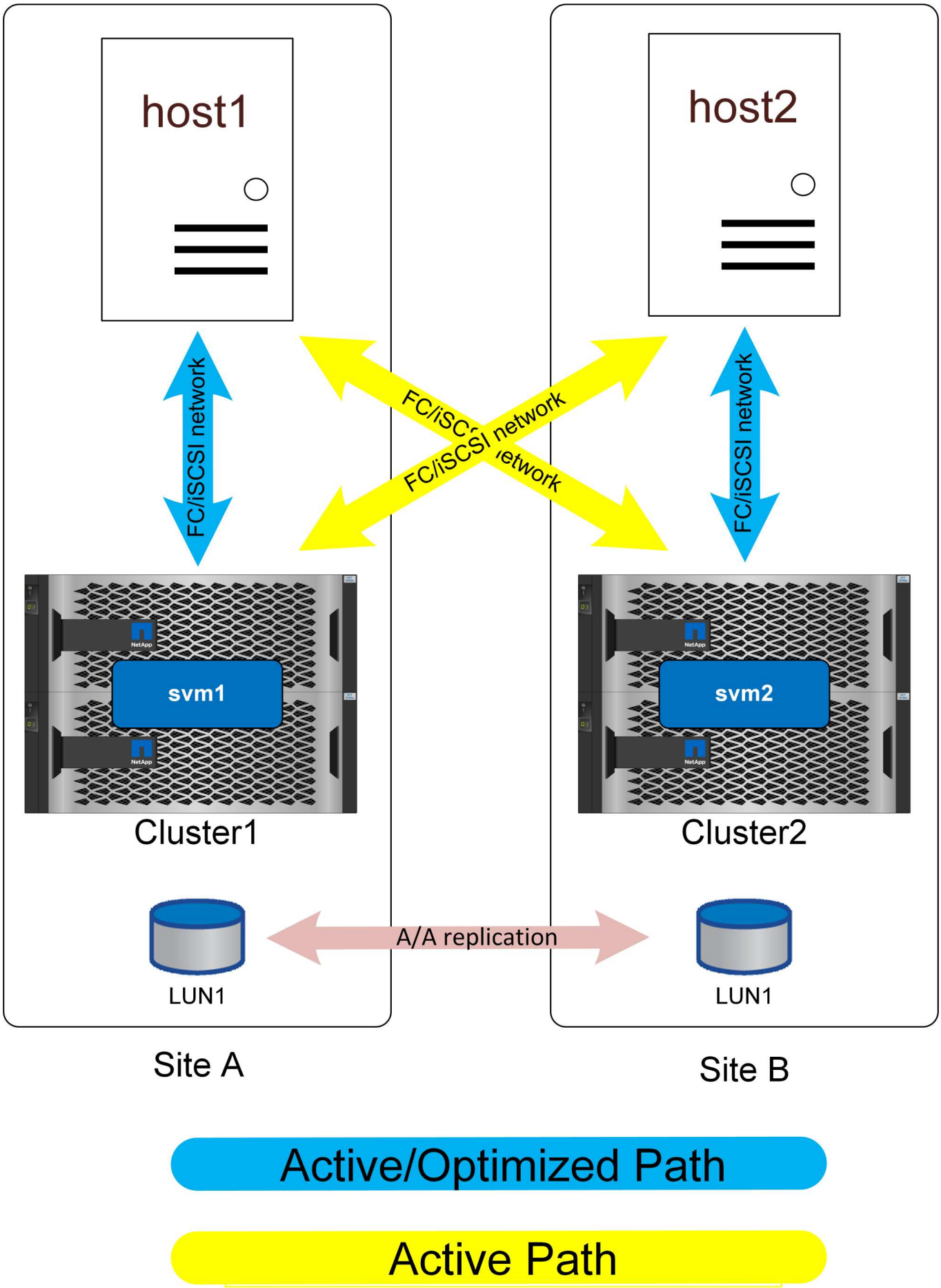
Ad esempio, un host del sito A potrebbe essere configurato in modo da accedere a una LUN protetta dalla sincronizzazione attiva di SnapMirror e, poiché la SAN è estesa tra i siti, i percorsi sono disponibili per tale LUN utilizzando lo storage sul sito A o lo storage sul sito B.

Senza impostazioni di prossimità, l'host utilizzerà entrambi i sistemi storage allo stesso modo perché entrambi i sistemi storage pubblicizzeranno i percorsi attivi/ottimizzati. Se la latenza SAN e/o la larghezza di banda tra i siti è limitata, questa operazione potrebbe non essere disattivabile e potrebbe essere necessario assicurarsi

che durante il normale funzionamento ogni host utilizzi preferenzialmente i percorsi verso il sistema di storage locale. Viene configurato aggiungendo l'ID WWN/iSCSI dell'host al cluster locale come host prossimale. Questa operazione può essere eseguita dalla CLI o da SystemManager.

AFF

Con un sistema AFF, i percorsi vengono visualizzati come mostrato di seguito quando è stata configurata la prossimità dell'host.


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

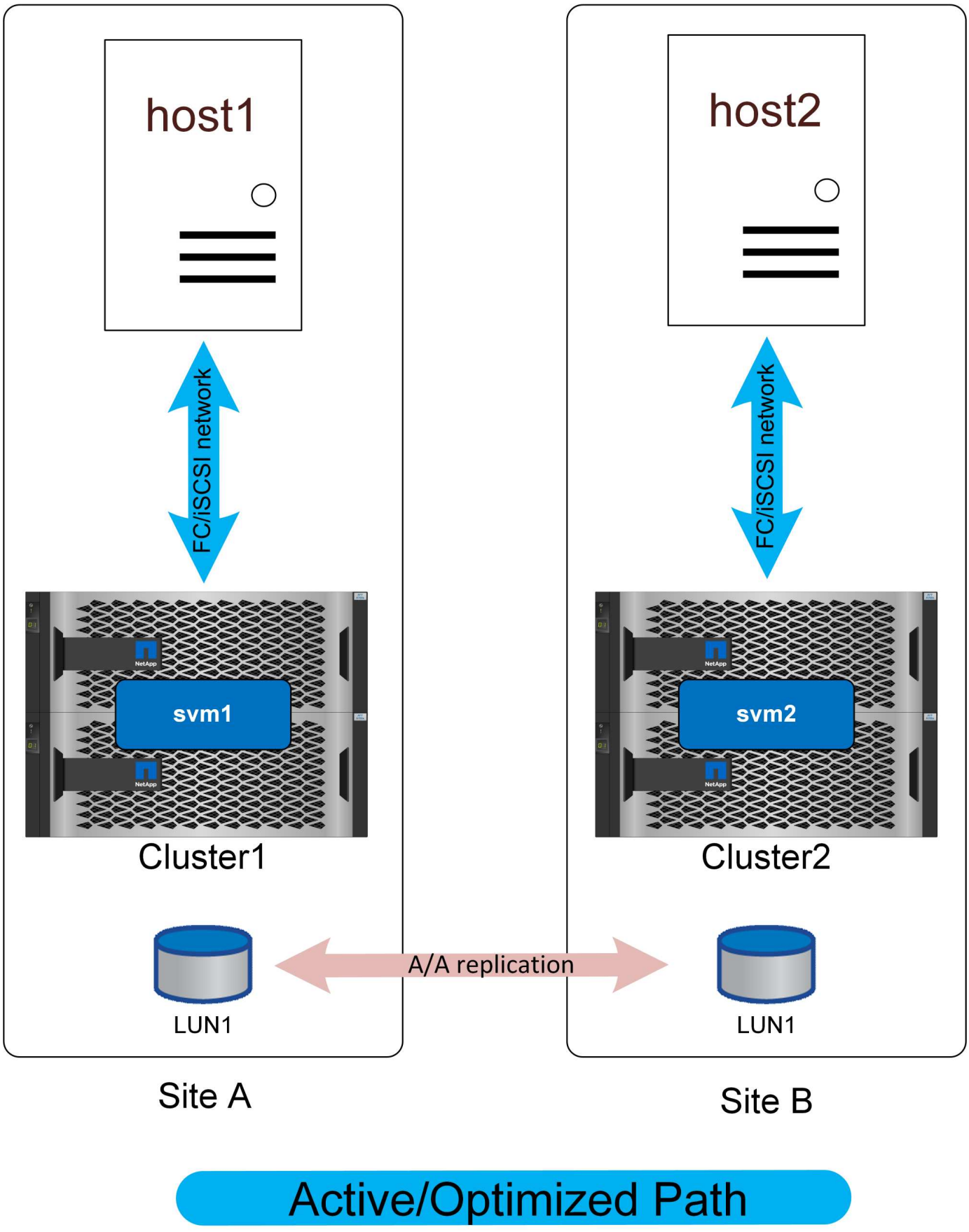
Durante le normali operazioni, tutti i io sono i/o locali. Le letture e le scritture sono gestite dallo storage array locale. Gli io in scrittura, naturalmente, dovranno anche essere replicati dal controller locale sul sistema remoto prima di essere riconosciuti, ma tutti gli io in lettura saranno serviti localmente e non subiranno una latenza aggiuntiva attraverso il collegamento SAN tra i siti.

L'unica volta in cui verranno utilizzati i percorsi non ottimizzati è quando tutti i percorsi attivi/ottimizzati vengono persi. Ad esempio, se l'intero array sul sito A perde energia, gli host sul sito A sarebbero comunque in grado di accedere ai percorsi dell'array sul sito B e di rimanere quindi operativi, anche se sperimenterebbero una latenza più elevata.

Esistono percorsi ridondanti attraverso il cluster locale non mostrati in questi diagrammi per semplicità. I sistemi di storage ONTAP sono ad alta disponibilità, pertanto un guasto a un controller non dovrebbe causare guasti nel sito. Ciò dovrebbe comportare solo una modifica dei percorsi locali utilizzati nel sito interessato.

ASA

I sistemi NetApp ASA offrono multipathing Active-Active su tutti i percorsi di un cluster. Questo vale anche per le configurazioni SM-AS.



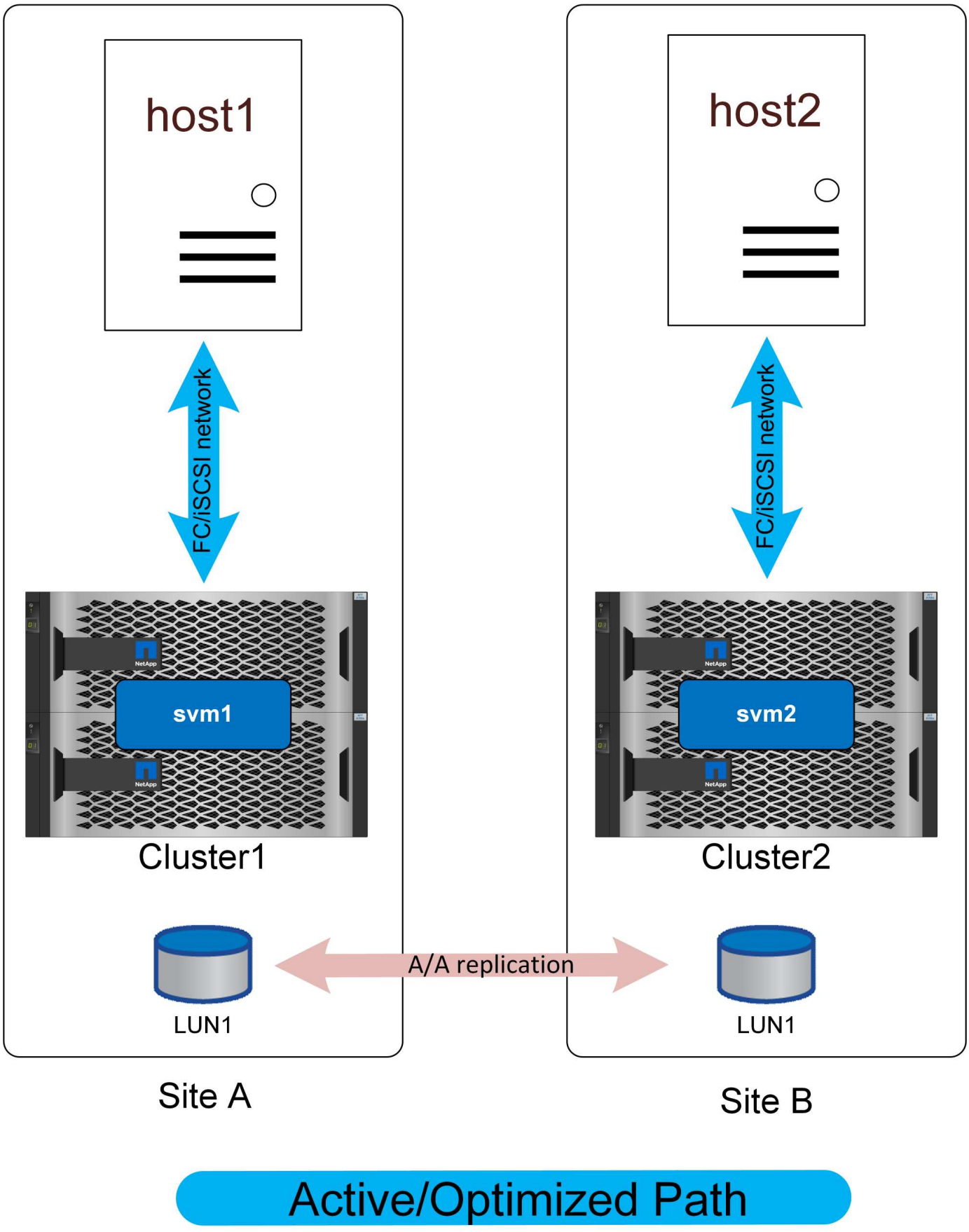
Una configurazione ASA con accesso non uniforme funzionerebbe in gran parte allo stesso modo di AFF. Con un accesso uniforme, io attraverserebbe la WAN. Ciò può essere o non può essere desiderabile.

Se i due siti fossero a una distanza di 100 metri con connettività in fibra non dovrebbe esserci una latenza aggiuntiva rilevabile che attraversa la WAN, ma se i siti fossero a lunga distanza gli uni dagli altri, le performance in lettura risulterebbero compromesse su entrambi i siti. Al contrario, con AFF questi percorsi WAN sarebbero utilizzati solo se non ci fossero percorsi locali disponibili e le performance quotidiane sarebbero migliori, perché tutti i io sarebbero io locali. ASA con rete di accesso non uniforme sarebbe un'opzione per ottenere i vantaggi in termini di costi e funzionalità di ASA senza incorrere in penalizzazioni per l'accesso alla latenza tra siti.

ASA con SM-as in una configurazione a bassa latenza offre due benefici interessanti. In primo luogo, essenzialmente **raddoppia** le prestazioni per ogni singolo host perché io può essere gestito dal doppio dei controller utilizzando il doppio dei percorsi. In secondo luogo, in un ambiente a sito singolo offre una disponibilità estrema, perché è possibile perdere un intero sistema storage senza interrompere l'accesso degli host.

Accesso non uniforme

Una rete di accesso non uniforme significa che ogni host ha solo accesso alle porte sul sistema di storage locale. LA SAN non è estesa a più siti (o presenta errori nei domini dello stesso sito).



Il vantaggio principale di questo approccio è la semplicità DELLE SAN, eliminando l'esigenza di stretching di una SAN via rete. Alcuni clienti non dispongono di connettività a latenza sufficientemente bassa tra i siti o non

dispongono dell'infrastruttura per il tunnel del traffico FC SAN su una rete intersito.

Lo svantaggio legato all'accesso non uniforme è che alcuni scenari di errore, inclusa la perdita del collegamento di replica, provocheranno la perdita dell'accesso allo storage da parte di alcuni host. In caso di interruzione della connettività dello storage locale, le applicazioni eseguite come istanze singole, come ad esempio i database non in cluster, eseguiti in maniera intrinseca solo su un singolo host in uno qualsiasi dei supporti di montaggio, si guasterebbero. I dati sarebbero comunque protetti, ma il server di database non avrebbe più accesso. Dovrebbe essere riavviato su un sito remoto, preferibilmente tramite un processo automatizzato. Ad esempio, VMware ha è in grado di rilevare una situazione di tutti i percorsi verso l'esterno su un server e di riavviare una macchina virtuale su un altro server in cui sono disponibili i percorsi.

Al contrario, un'applicazione in cluster come Oracle RAC può offrire un servizio disponibile contemporaneamente in due siti diversi. Perdere un sito non significa perdere il servizio dell'applicazione nel suo complesso. Le istanze sono ancora disponibili e in esecuzione nel sito sopravvissuto.

In molti casi, l'overhead della latenza aggiuntivo di un'applicazione che accede allo storage attraverso un collegamento da sito a sito sarebbe inaccettabile. Ciò significa che la migliore disponibilità di una rete uniforme è minima, poiché la perdita di storage su un sito comporterebbe comunque la necessità di arrestare i servizi sul sito in cui si è verificato l'errore.

> Esistono percorsi ridondanti attraverso il cluster locale non mostrati in questi diagrammi per semplicità. I sistemi di storage ONTAP sono ad alta disponibilità, pertanto un guasto a un controller non dovrebbe causare guasti nel sito. Ciò dovrebbe comportare solo una modifica dei percorsi locali utilizzati nel sito interessato.

Configurazioni Oracle

Panoramica

L'utilizzo della sincronizzazione attiva di SnapMirror non aggiunge o modifica necessariamente le procedure consigliate per il funzionamento di un database.

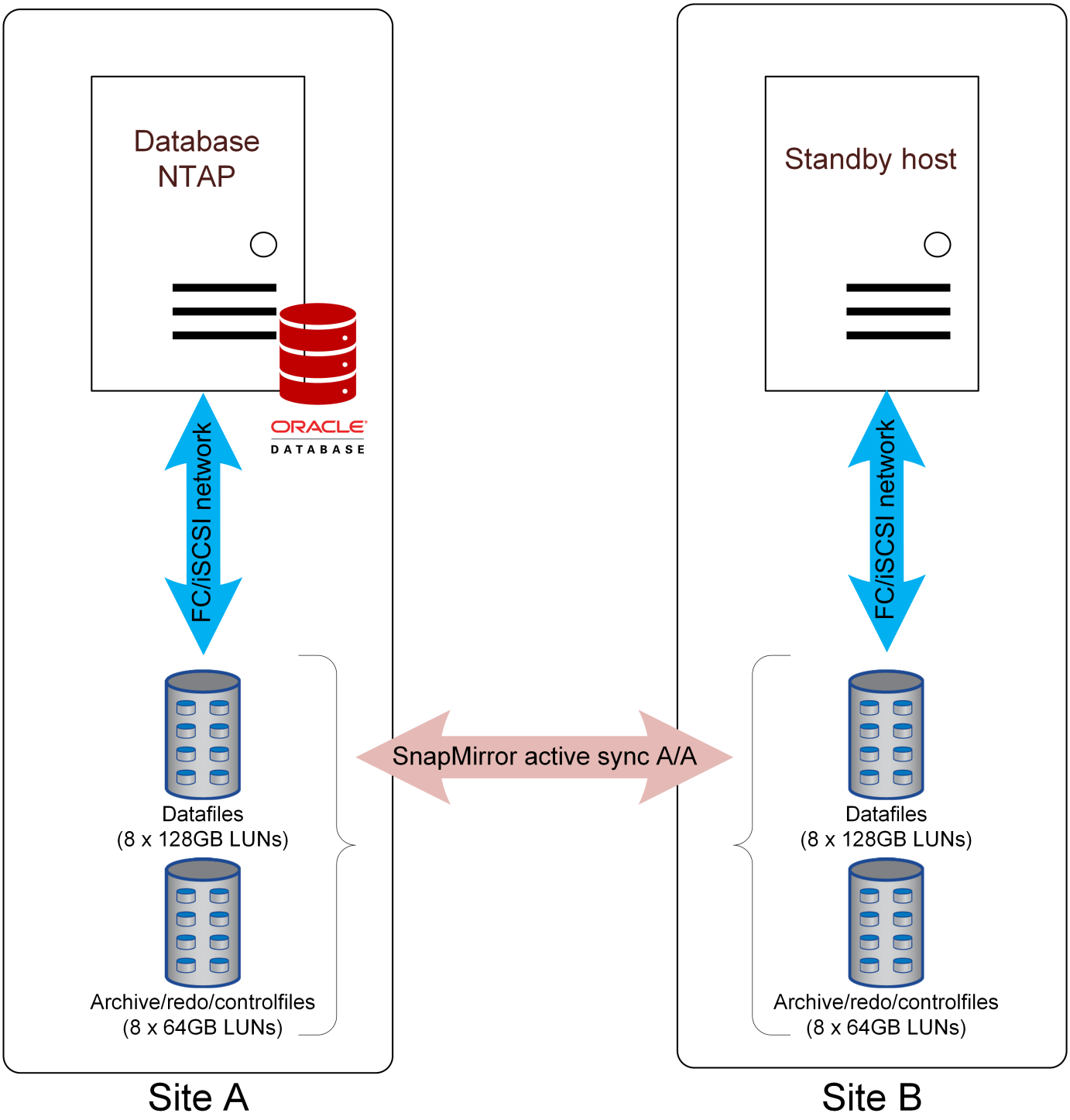
La migliore architettura dipende dai requisiti di business. Ad esempio, se l'obiettivo è ottenere una protezione RPO=0:1 contro la perdita di dati, ma l'RTO non è più così semplice, utilizzare i database Oracle Single Instance e replicare le LUN con SM, potrebbe essere sufficiente così come meno costoso da un approccio basato su licenze Oracle. Un guasto del sito remoto non interromperebbe le operazioni e, di conseguenza, la perdita del sito primario porterebbe all'utilizzo di LUN online e pronte per l'uso nel sito rimasto.

Se l'RTO fosse più rigoroso, l'automazione Active-passive di base tramite script o cluster come Pacemaker o Ansible migliorerebbe i tempi di failover. Ad esempio, VMware ha può essere configurato in modo da rilevare un guasto della VM nel sito primario e rendere attiva la VM nel sito remoto.

Infine, per un failover estremamente rapido, Oracle RAC può essere implementato su più siti. L'RTO sarebbe praticamente pari a zero perché il database sarebbe sempre online e disponibile su entrambi i siti.

Istanza singola Oracle

Gli esempi illustrati di seguito mostrano alcune delle numerose opzioni per la distribuzione dei database Oracle Single Instance con la replica di sincronizzazione attiva SnapMirror.



Failover con un sistema operativo preconfigurato

La sincronizzazione attiva SnapMirror fornisce una copia sincrona dei dati nel sito di disaster recovery, ma la loro disponibilità richiede un sistema operativo e le applicazioni associate. L'automazione di base può migliorare notevolmente il tempo di failover dell'ambiente complessivo. I prodotti Clusterware come Pacemaker vengono spesso utilizzati per creare un cluster in tutti i siti, e in molti casi il processo di failover può essere guidato con semplici script.

In caso di perdita dei nodi primari, il clusterware (o gli script) porterà i database online nel sito alternativo. Un'opzione è creare server di standby preconfigurati per le risorse SAN che compongono il database. Se il sito primario non funziona, il clusterware o l'alternativa con script esegue una sequenza di azioni simile alle

seguenti:

1. Rileva i guasti del sito primario
2. Rilevamento di LUN FC o iSCSI
3. Montaggio di file system e/o montaggio di gruppi di dischi ASM
4. Avvio del database

Il requisito principale di questo approccio è rappresentato da un sistema operativo in esecuzione sul sito remoto. Deve essere preconfigurato con i file binari di Oracle, il che significa anche che attività come l'applicazione delle patch Oracle devono essere eseguite sul sito primario e di standby. In alternativa, è possibile eseguire il mirroring dei file binari di Oracle nel sito remoto e montarli se viene dichiarato un disastro.

La procedura di attivazione effettiva è semplice. Comandi come il rilevamento delle LUN richiedono solo pochi comandi per ogni porta FC. Il montaggio del file system non è altro che un `mount` comando e sia i database che ASM possono essere avviati e arrestati dalla CLI con un unico comando.

Failover con un sistema operativo virtualizzato

Il failover degli ambienti di database può essere esteso per includere il sistema operativo stesso. In teoria, questo failover può essere eseguito con le LUN di avvio, ma nella maggior parte dei casi con un sistema operativo virtualizzato. La procedura è simile ai seguenti passaggi:

1. Rileva i guasti del sito primario
2. Montaggio dei datastore che ospitano le macchine virtuali del server di database
3. Avvio delle macchine virtuali
4. Avviare i database manualmente o configurare le macchine virtuali per avviare automaticamente i database.

Ad esempio, un cluster ESX può estendersi su diversi siti. In caso di disastro, dopo lo switchover, è possibile portare online le macchine virtuali nel sito di disaster recovery.

Protezione dai guasti dello storage

Il diagramma precedente mostra l'utilizzo di "accesso non uniforme", in cui la SAN non è estesa tra i siti. Questa configurazione potrebbe essere più semplice e, in alcuni casi, potrebbe essere l'unica opzione data le attuali funzionalità SAN, ma significa anche che un guasto del sistema di storage primario causerebbe un'interruzione del database fino a quando l'applicazione non è stata sottoposta a failover.

Per una maggiore resilienza, la soluzione potrebbe essere implementata con "accesso uniforme". Ciò consentirebbe alle applicazioni di continuare a funzionare utilizzando i percorsi pubblicizzati dal sito opposto.

Oracle Extended RAC

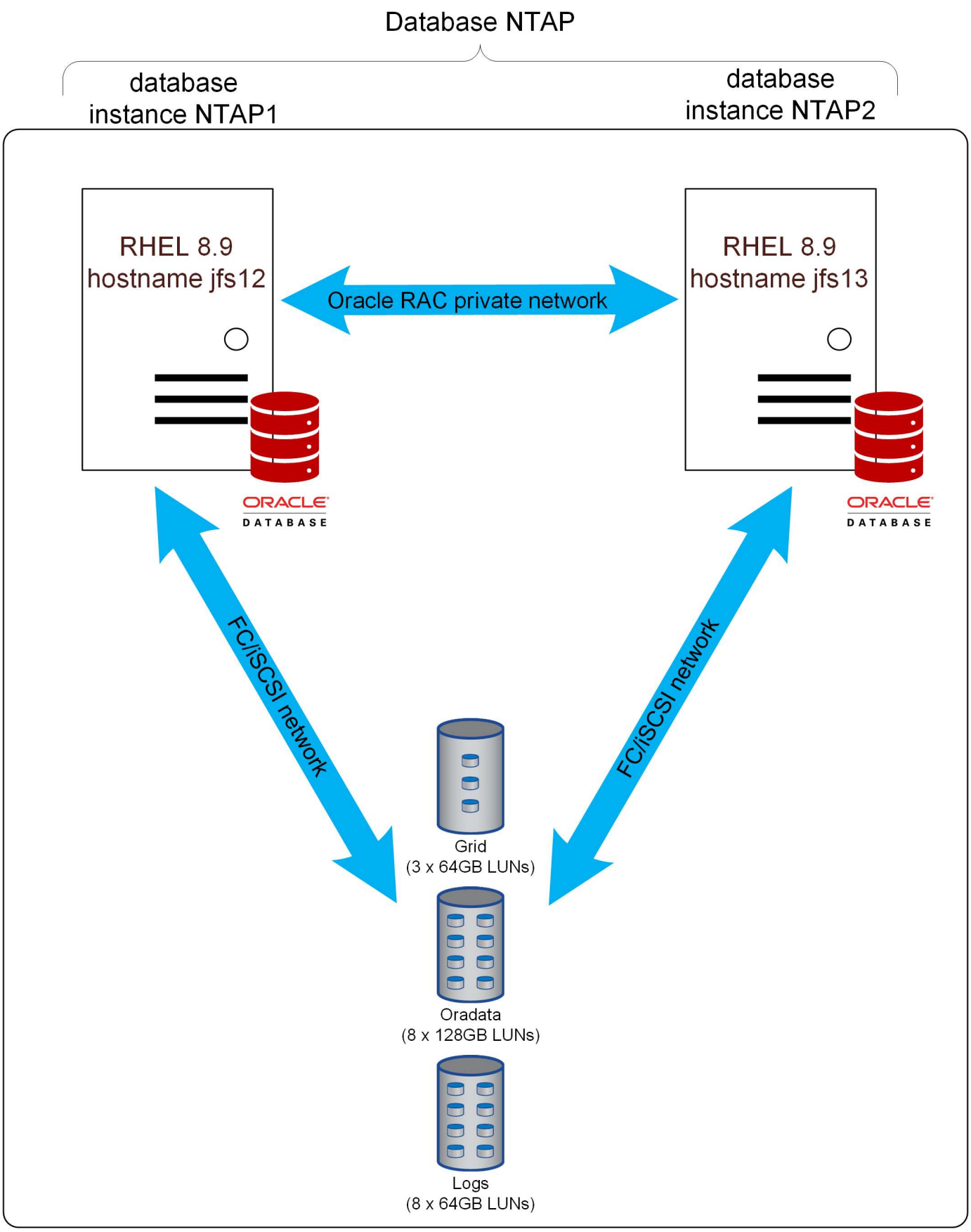
Molti clienti ottimizzano il proprio RTO estendendo un cluster Oracle RAC tra i vari siti, ottenendo una configurazione completamente Active-Active. La progettazione complessiva diventa più complicata perché deve includere la gestione del quorum di Oracle RAC.

Il cluster RAC esteso tradizionale si basava sul mirroring ASM per garantire la protezione dei dati. Questo approccio funziona, ma richiede anche molti passaggi di configurazione manuale e impone un overhead all'infrastruttura di rete. Al contrario, consentendo alla sincronizzazione attiva di SnapMirror di assumersi la

responsabilità della replica dei dati, la soluzione risulta notevolmente semplificata. Operazioni quali sincronizzazione, risincronizzazione dopo interruzioni, failover e gestione del quorum sono più semplici, inoltre la SAN non deve essere distribuita tra i siti, semplificando la progettazione e la gestione della SAN.

Replica

La chiave per comprendere le funzionalità RAC nella sincronizzazione attiva di SnapMirror è la visualizzazione dello storage come singolo set di LUN che si trovano sullo storage con mirroring. Ad esempio:



Nessuna copia primaria o copia speculare. A livello logico, esiste una sola copia di ogni LUN e tale LUN è disponibile nei percorsi SAN che si trovano su due sistemi di storage differenti. Dal punto di vista dell'host, non ci sono failover dello storage e modifiche del percorso. Diversi eventi di errore potrebbero causare la perdita di

determinati percorsi verso la LUN, mentre gli altri percorsi rimangono online. La sincronizzazione attiva di SnapMirror garantisce la disponibilità degli stessi dati su tutti i percorsi operativi.

Configurazione dello storage

In questa configurazione di esempio, i dischi ASM sono configurati come in qualsiasi configurazione RAC a sito singolo sullo storage Enterprise. Poiché il sistema di storage garantisce la protezione dei dati, viene utilizzata la ridondanza esterna ASM.

Accesso uniforme e non privo di informazioni

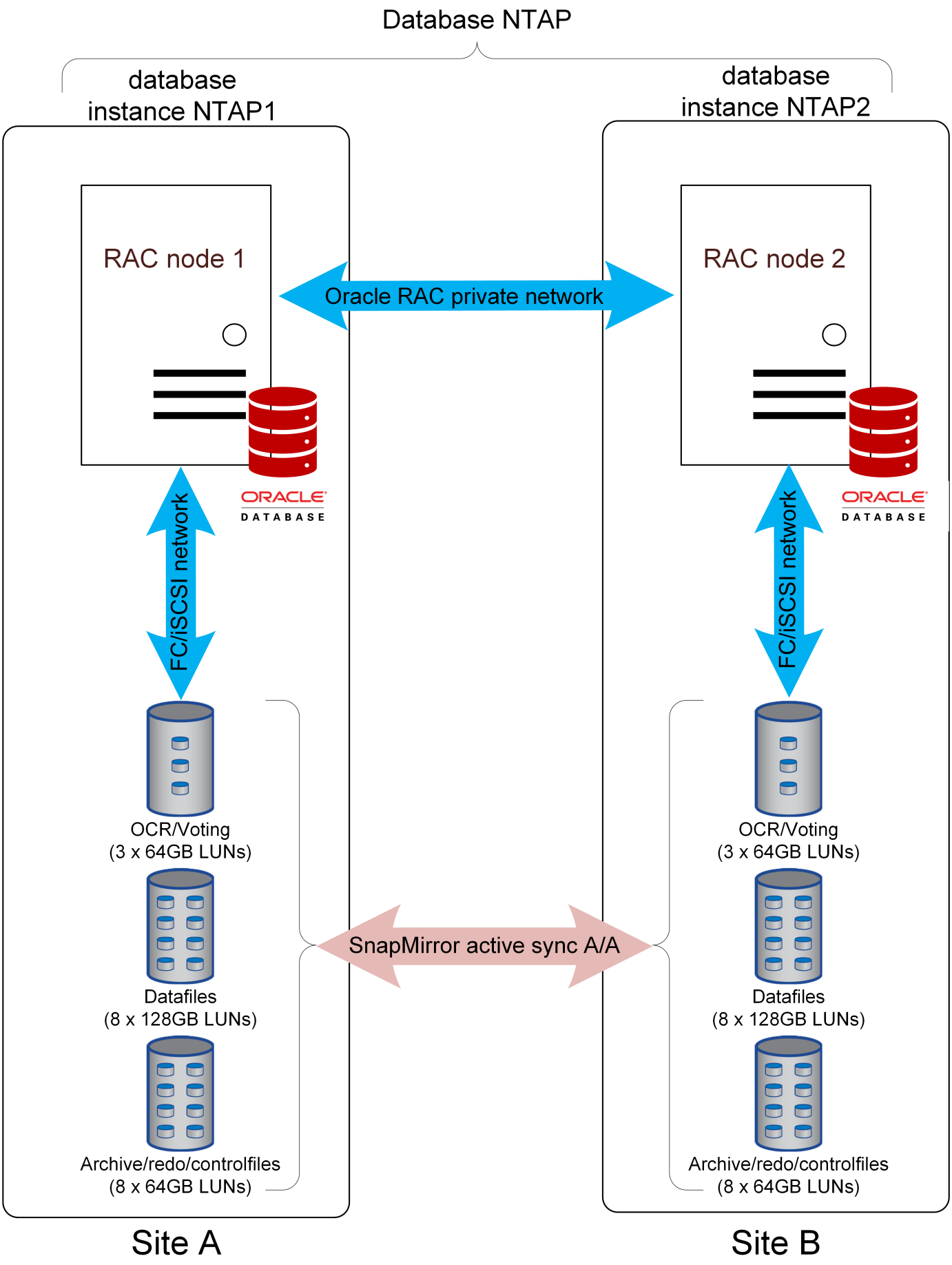
La considerazione più importante con Oracle RAC sulla sincronizzazione attiva di SnapMirror è se utilizzare un accesso uniforme o non uniforme.

Un accesso uniforme significa che ogni host può vedere i percorsi su entrambi i cluster. Accesso non uniforme significa che gli host possono vedere solo i percorsi verso il cluster locale.

Nessuna delle due opzioni è specificamente consigliata o scoraggiata. Alcuni clienti dispongono prontamente di fibra ottica spenta per la connessione ai siti, altri non dispongono di tale connettività oppure la loro infrastruttura SAN non supporta un ISL a lunga distanza.

Accesso non uniforme

L’accesso non uniforme è più semplice da configurare dal punto di vista della SAN.


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

L’aspetto negativo principale "accesso non uniforme"dell’approccio è rappresentato dalla perdita della connettività ONTAP sito-sito o dalla perdita di un sistema di storage che causa la perdita delle istanze di database in un singolo sito. Questo ovviamente non è desiderabile, ma potrebbe essere un rischio accettabile in cambio di una configurazione SAN più semplice.

Accesso uniforme

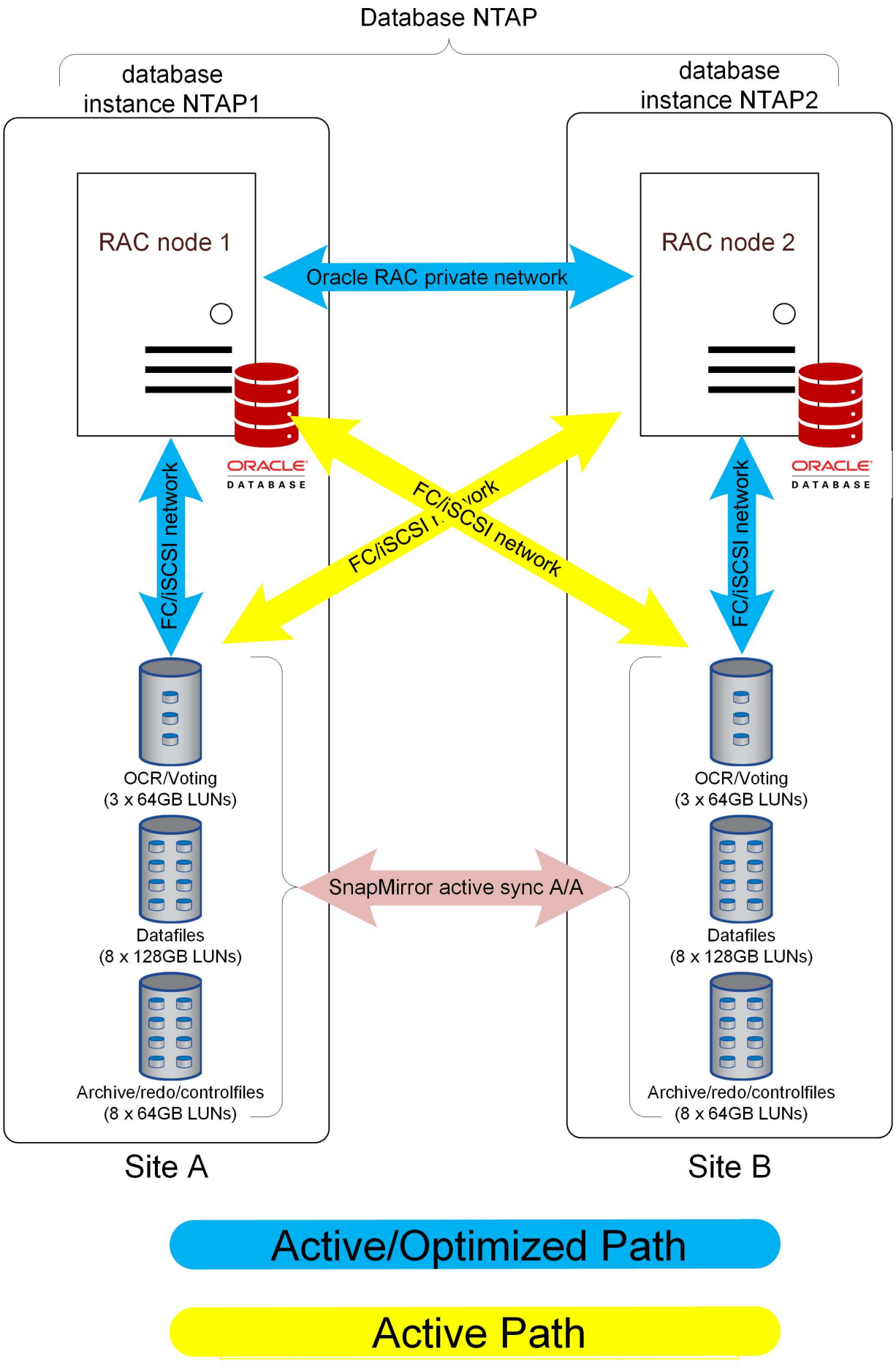
Un accesso uniforme richiede l’estensione della SAN tra i siti. Il vantaggio principale consiste nel fatto che la perdita di un sistema di storage non provocherà la perdita di un’istanza del database. Al contrario, si otterrebbe una modifica multipathing in cui i percorsi sono attualmente in uso.

Esistono diversi modi per configurare l’accesso non uniforme.

> Nei diagrammi di seguito sono presenti anche percorsi attivi ma non ottimizzati che sarebbero utilizzati durante semplici guasti del controller, ma tali percorsi non sono mostrati nell’interesse di semplificare i diagrammi.

AFF con impostazioni di prossimità

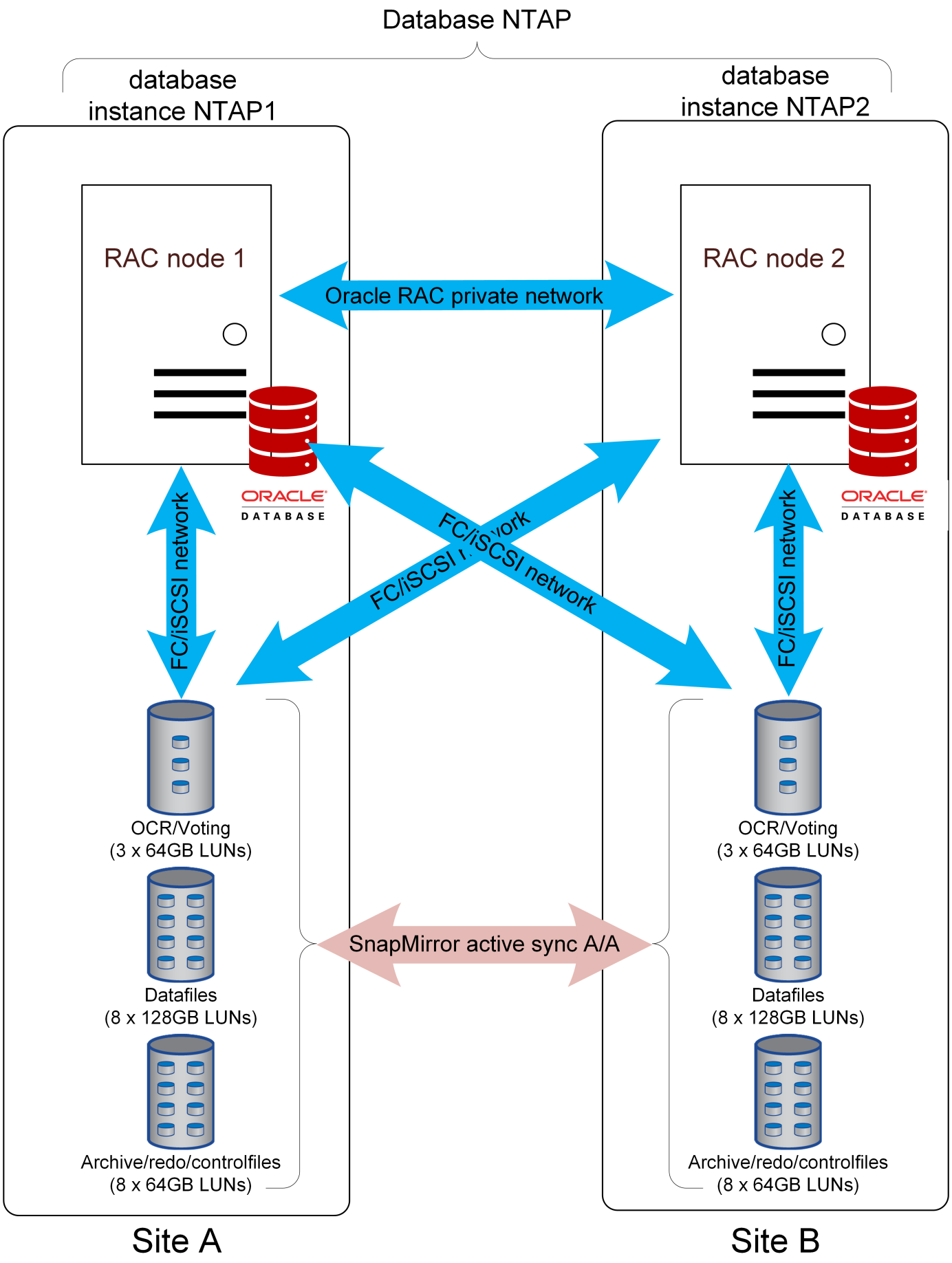
In presenza di una latenza significativa tra i siti, è possibile configurare i sistemi AFF con le impostazioni di prossimità dell’host. In questo modo, ciascun sistema storage può conoscere gli host locali e remoti e assegnare in maniera appropriata le priorità del percorso.


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

Durante il normale funzionamento, ogni istanza Oracle utilizzerebbe preferenzialmente i percorsi locali attivi/ottimizzati. Il risultato è che tutte le letture saranno gestite dalla copia locale dei blocchi. In questo modo, si ottiene la minore latenza possibile. L'io in scrittura viene analogamente inviato ai percorsi al controller locale. L'io deve ancora essere replicato prima di essere riconosciuto e quindi sarebbe comunque soggetto alla latenza aggiuntiva che attraversa la rete site-to-site, ma ciò non può essere evitato in una soluzione di replica sincrona.

ASA / AFF senza impostazioni di prossimità

In assenza di una latenza significativa tra i siti, è possibile configurare i sistemi AFF senza le impostazioni di prossimità dell'host oppure utilizzare ASA.


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

Ciascun host sarà in grado di utilizzare tutti i percorsi operativi su entrambi i sistemi storage. Questo consente di migliorare potenzialmente le prestazioni consentendo a ciascun host di sfruttare il potenziale in termini di prestazioni di due cluster, non di uno solo.

Con ASA, non solo tutti i percorsi verso entrambi i cluster sono considerati attivi e ottimizzati, ma anche i percorsi sui partner controller sarebbero attivi. Ne risulterebbero percorsi SAN all-Active sull'intero cluster, in qualsiasi momento.

> I sistemi ASA possono anche essere utilizzati in una configurazione di accesso non uniforme. Poiché non esistono percorsi tra siti, non vi sarebbe alcun impatto sulle performance risultanti dall'io che attraversa l'ISL.

RAC tiebreaker

Sebbene il RAC esteso che utilizza la sincronizzazione attiva di SnapMirror sia un'architettura simmetrica rispetto all'io, esiste un'eccezione connessa alla gestione split-brain.

Cosa succede se il collegamento di replica viene perso e nessuno dei due siti ha un quorum? Che cosa dovrebbe accadere? Questa domanda è valida sia per Oracle RAC che per il comportamento di ONTAP. Se non è possibile replicare le modifiche tra i siti e si desidera riprendere le operazioni, uno dei siti dovrà sopravvivere e l'altro sito dovrà diventare non disponibile.

La "Mediatore ONTAP" soddisfa questo requisito al livello ONTAP. Esistono diverse opzioni per RAC Tiebreaking.

Gli Oracle Tiebreaker

Il metodo migliore per gestire i rischi di Oracle RAC split-brain consiste nell'utilizzare un numero dispari di nodi RAC, preferibilmente utilizzando un Tiebreaker a 3rd siti. Se un sito 3rd non è disponibile, l'istanza di Tiebreaker potrebbe essere posizionata in un sito dei due siti, designando effettivamente un sito di sopravvivenza preferito.

Oracle e css_Critical

Con un numero pari di nodi, il comportamento predefinito di Oracle RAC è che uno dei nodi nel cluster sarà considerato più importante degli altri nodi. Il sito con tale nodo con priorità più alta sopravviverà all'isolamento del sito, mentre i nodi sull'altro sito verranno eliminati. La priorità si basa su più fattori, ma è anche possibile controllare questo comportamento utilizzando l' `css_critical`impostazione.

Nell'"esempio"architettura, i nomi host per i nodi RAC sono jfs12 e jfs13. Di seguito sono riportate le impostazioni correnti di `css_critical`:

```
[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.
```

Se si desidera che il sito con jfs12 sia il sito preferito, impostare questo valore su sì su un nodo del sito A e riavviare i servizi.

```
[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.
```

Scenari di errore

Panoramica

La pianificazione di un'architettura completa dell'applicazione SnapMirror Active Sync richiede la comprensione del modo in cui SM-AS risponderà in vari scenari di failover pianificati e non pianificati.

Per gli esempi seguenti, si supponga che il sito A sia configurato come sito preferito.

Interruzione della connettività di replica

Se la replica SM-AS viene interrotta, non è possibile completare la scrittura io perché sarebbe impossibile per un cluster replicare le modifiche al sito opposto.

Sito A (sito preferito)

Il risultato dell'errore del collegamento di replica sul sito preferito sarà una pausa di circa 15 secondi nell'elaborazione io in scrittura, poiché ONTAP ritenta le operazioni di scrittura replicate prima di determinare che il collegamento di replica è veramente irraggiungibile. Trascorsi i 15 secondi, il sistema del sito A riprende l'elaborazione io in lettura e scrittura. I percorsi SAN non vengono modificati e i LUN rimangono online.

Sito B

Poiché il sito B non è il sito preferito di sincronizzazione attiva SnapMirror, i relativi percorsi LUN non saranno più disponibili dopo circa 15 secondi.

Errore del sistema di storage

Il risultato di un errore del sistema di storage è quasi identico al risultato della perdita del collegamento di replica. Il sito sopravvissuto dovrebbe subire una pausa io di circa 15 secondi. Trascorso questo periodo di 15 secondi, io riprenderà sul sito come di consueto.

Perdita del mediatore

Il servizio di mediazione non controlla direttamente le operazioni di storage. Funziona come un percorso di controllo alternativo tra cluster. Esiste principalmente per automatizzare il failover senza il rischio di uno scenario split-brain. Durante l'utilizzo normale, ogni cluster replica le modifiche al partner e pertanto ogni cluster può verificare che il cluster partner sia online e fornisca i dati. Se il collegamento di replica non è riuscito, la replica viene interrotta.

Il motivo per cui è necessario un mediatore per il failover automatizzato sicuro è perché altrimenti sarebbe impossibile per un cluster di storage determinare se la perdita di comunicazione bidirezionale fosse il risultato di un'interruzione della rete o di un errore effettivo dello storage.

Il mediatore fornisce un percorso alternativo per ciascun cluster per verificare lo stato di salute del partner. Gli scenari sono i seguenti:

- Se un cluster può contattare direttamente il partner, i servizi di replica sono operativi. Non è richiesta alcuna azione.
- Se un sito preferito non può contattare direttamente il proprio partner o tramite il mediatore, presuppone che il partner sia effettivamente non disponibile oppure è stato isolato e ha portato i percorsi LUN offline. Il sito preferito procede quindi al rilascio dello stato RPO=0 e continua l'elaborazione dell'io in lettura e in scrittura.
- Se un sito non preferito non può contattare direttamente il proprio partner, ma può contattarlo tramite il mediatore, prenderà i suoi percorsi offline e attenderà il ritorno della connessione di replica.
- Se un sito non preferito non può contattare direttamente il proprio partner o tramite un mediatore operativo, supporterà che il partner sia effettivamente non disponibile, oppure che sia stato isolato e che abbia portato i percorsi LUN offline. Il sito non preferito procede quindi al rilascio dello stato RPO=0 e continua l'elaborazione dell'i/o in lettura e scrittura. Assumerà il ruolo dell'origine della replica e diventerà il nuovo sito preferito.

Se il mediatore non è completamente disponibile:

- In caso di guasto dei servizi di replica per qualsiasi motivo, incluso un guasto del sito o del sistema storage non preferito, il sito preferito rilascerà lo stato RPO=0 e riprenderà l'elaborazione i/o in lettura e scrittura. Il sito non preferito prenderà i suoi percorsi offline.
- Il guasto del sito preferito causerà un'interruzione poiché il sito non preferito non sarà in grado di verificare che il sito opposto sia effettivamente offline e quindi non sarebbe sicuro per il sito non preferito riprendere i servizi.

Ripristino dei servizi in corso

Dopo aver risolto un errore, come il ripristino della connettività da sito a sito o l'accensione di un sistema guasto, gli endpoint di sincronizzazione attivi di SnapMirror rilevano automaticamente la presenza di una relazione di replica difettosa e la riportano allo stato RPO=0. Una volta ristabilita la replica sincrona, i percorsi non riusciti torneranno in linea.

In molti casi, le applicazioni in cluster rilevano automaticamente la restituzione dei percorsi non riusciti e tali applicazioni tornano online. In altri casi, potrebbe essere necessaria una scansione SAN a livello di host

oppure potrebbe essere necessario riportare le applicazioni online manualmente. Dipende dall'applicazione e dal modo in cui è configurata, e in generale tali attività possono essere facilmente automatizzate. ONTAP si sta auto-riparando e non deve richiedere alcun intervento da parte dell'utente per riprendere le operazioni di storage RPO = 0 KB.

Failover manuale

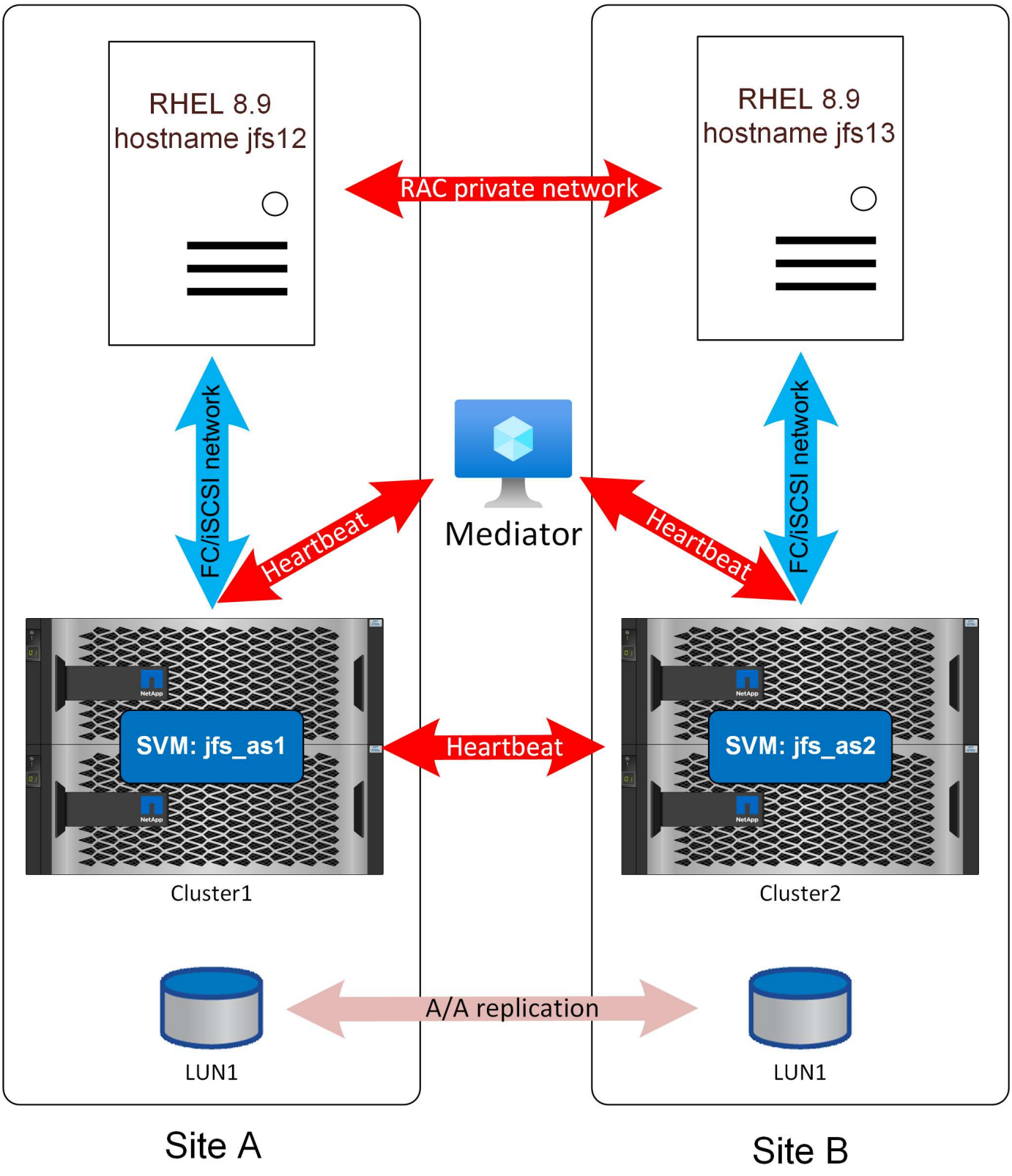
La modifica del sito preferito richiede un'operazione semplice. I/o si fermeranno per un secondo o due come autorità sugli switch del comportamento di replica tra i cluster, ma in caso contrario i/o non vengono influenzati.

Architettura di esempio

Gli esempi dettagliati di guasto illustrati in questa sezione si basano sull'architettura illustrata di seguito.

> Questa è solo una delle numerose opzioni per i database Oracle su SnapMirror Active Sync. Questo disegno è stato scelto perché illustra alcuni degli scenari più complicati.

In questa configurazione, si supponga che il sito A sia impostato su "sito preferito".



Errore di interconnessione RAC

La perdita del collegamento di replica di Oracle RAC produrrà un risultato simile alla perdita della connettività SnapMirror, tuttavia i timeout saranno più brevi per impostazione predefinita. Con le impostazioni predefinite, un nodo Oracle RAC attenderà 200 secondi

dopo la perdita della connettività di storage prima dell'eliminazione, ma attenderà solo 30 secondi dopo la perdita dell'heartbeat della rete RAC.

I messaggi CRS sono simili a quelli illustrati di seguito. È possibile visualizzare il timeout di 30 secondi. Poiché css_Critical è stato impostato su jfs12, situato sul sito A, questo sarà il sito per sopravvivere e jfs13 sul sito B sarà espulso.

```
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'.
```

Errore di comunicazione SnapMirror

Se il collegamento di replica sincrona attiva di SnapMirror, l'io in scrittura non può essere completato perché sarebbe impossibile per un cluster replicare le modifiche al sito opposto.

Sito A

Il risultato sul sito A di un errore del collegamento di replica sarà una pausa di circa 15 secondi nell'elaborazione io in scrittura, poiché ONTAP tenta di replicare le scritture prima di determinare che il collegamento di replica è veramente inutilizzabile. Trascorsi i 15 secondi, il cluster ONTAP sul sito A riprende l'elaborazione i/o in lettura e scrittura. I percorsi SAN non vengono modificati e i LUN rimangono online.

Sito B

Poiché il sito B non è il sito preferito di sincronizzazione attiva SnapMirror, i relativi percorsi LUN non saranno più disponibili dopo circa 15 secondi.

Il collegamento di replica è stato tagliato alla data e ora 15:19:44. Il primo avviso da Oracle RAC arriva 100 secondi dopo l'avvicinarsi del timeout di 200 secondi (controllato dal disktimeout del parametro di Oracle RAC).

```
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.
```

Una volta raggiunto il timeout del disco di voting di 200 secondi, il nodo Oracle RAC verrà rimosso dal cluster e riavviato.

Errore interconnettività di rete totale

Se il collegamento di replica tra i siti viene perso completamente, la sincronizzazione attiva di SnapMirror e la connettività Oracle RAC verranno interrotte.

Il rilevamento split-brain di Oracle RAC dipende dall'heartbeat dello storage di Oracle RAC. Se la perdita della connettività da sito a sito determina la perdita simultanea dei servizi di replica dello storage e heartbeat della rete RAC, i siti RAC non saranno in grado di comunicare tra siti tramite RAC Interconnect o i dischi di voto RAC. Il risultato in un insieme di nodi pari-numerati può essere l'esclusione di entrambi i siti nelle impostazioni di default. Il comportamento esatto dipende dalla sequenza degli eventi e dalla tempistica dei sondaggi della rete RAC e degli heartbeat del disco.

È possibile risolvere il rischio di un'interruzione dei servizi dei siti 2 in due modi. In primo luogo, è possibile utilizzare una "tiebreaker" configurazione.

Se non è disponibile un sito 3rd, questo rischio può essere risolto regolando il parametro misscount sul cluster RAC. Per impostazione predefinita, il timeout heartbeat della rete RAC è di 30 secondi. Normalmente viene utilizzato dal RAC per identificare i nodi RAC guasti e rimuoverli dal cluster. Ha anche una connessione al disco di voto heartbeat.

Se, ad esempio, il condotto che trasporta il traffico tra i siti per Oracle RAC e i servizi di replica dello storage viene tagliato da un retroescavatore, inizia il conto alla rovescia per gli errori di 30 secondi. Se il nodo del sito preferito RAC non riesce a ristabilire il contatto con il sito opposto entro 30 secondi e non può utilizzare i dischi di voto per confermare che il sito opposto non sia attivo entro la stessa finestra di 30 secondi, anche i nodi del sito preferito verranno eliminati. Il risultato è un'interruzione completa del database.

A seconda di quando si verifica il polling dell'errore scount, 30 secondi potrebbero non essere sufficienti per il timeout della sincronizzazione attiva SnapMirror e consentire allo storage sul sito preferito di riprendere i servizi prima della scadenza della finestra di 30 secondi. Questa finestra di 30 secondi può essere aumentata.

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

Questo valore consente al sistema di storage sul sito preferito di riprendere le operazioni prima della scadenza del timeout per il conteggio degli errori. Il risultato sarà quindi l'eliminazione solo dei nodi nel sito in cui sono stati rimossi i percorsi LUN. Esempio seguente:

```
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
```

Oracle Support sconsiglia vivamente di modificare i parametri misscount o disktimeout per risolvere i problemi di configurazione. La modifica di questi parametri può tuttavia essere garantita e inevitabile in molti casi, incluso l'avvio SAN, la virtualizzazione e le configurazioni di replica dello storage. Se, ad esempio, si sono verificati problemi di stabilità con una rete SAN o IP che hanno provocato estrazioni RAC, è necessario risolvere il problema sottostante e non caricare i valori di misscount o disktimeout. La modifica dei timeout per correggere gli errori di configurazione è mascherare un problema, non risolvere un problema. La modifica di

questi parametri per configurare correttamente un ambiente RAC in base agli aspetti di progettazione dell'infrastruttura sottostante è diversa ed è coerente con le istruzioni di supporto di Oracle. Con l'avvio SAN, è comune regolare misscount fino a 200 per far corrispondere disktimeout. Per ulteriori informazioni, vedere"questo link".

Guasto del sito

Il risultato di un errore del sistema di storage o della sede è quasi identico al risultato della perdita del collegamento di replica. Il sito sopravvissuto dovrebbe subire una pausa io di circa 15 secondi sulle scritture. Trascorso questo periodo di 15 secondi, io riprenderà sul sito come di consueto.

Se il problema riguarda solo il sistema storage, il nodo Oracle RAC del sito guasto perderà i servizi di storage e inserirà lo stesso conto alla rovescia di 200 secondi per il disktime, prima dell'eliminazione e del successivo riavvio.

```
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.
```

Lo stato del percorso SAN sul nodo RAC che ha perso i servizi storage è simile al seguente:

```
oradata7 (3600a0980383041334a3f55676c697347) dm-20 NETAPP,LUN C-Mode
size=128G features='3 queue_if_no_path pg_init_retries 50' hwhandler='1
alua' wp=rw
|-+- policy='service-time 0' prio=0 status=enabled
| `- 34:0:0:18 sdam 66:96  failed faulty running
`-+- policy='service-time 0' prio=0 status=enabled
  `- 33:0:0:18 sdaj 66:48  failed faulty running
```

L'host linux ha rilevato la perdita dei percorsi molto più velocemente di 200 secondi, ma dal punto di vista del database le connessioni client all'host sul sito guasto saranno ancora bloccate per 200 secondi sotto le impostazioni predefinite di Oracle RAC. Le operazioni complete del database riprenderanno solo dopo che l'eviction è stata completata.

Nel frattempo, il nodo RAC di Oracle sul sito opposto registrerà la perdita dell'altro nodo RAC. In caso contrario, continua a funzionare come di consueto.

```
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
```

Errore mediatore

Il servizio di mediazione non controlla direttamente le operazioni di storage. Funziona come un percorso di controllo alternativo tra cluster. Esiste principalmente per automatizzare il failover senza il rischio di uno scenario split-brain.

Durante l'utilizzo normale, ogni cluster replica le modifiche al partner e pertanto ogni cluster può verificare che il cluster partner sia online e fornisca i dati. Se il collegamento di replica non è riuscito, la replica viene interrotta.

Il motivo per cui è necessario un mediatore per operazioni automatizzate sicure è che altrimenti sarebbe

impossibile per i cluster di storage determinare se la perdita di comunicazione bidirezionale fosse il risultato di un'interruzione della rete o di un errore effettivo dello storage.

Il mediatore fornisce un percorso alternativo per ciascun cluster per verificare lo stato di salute del partner. Gli scenari sono i seguenti:

- Se un cluster può contattare direttamente il partner, i servizi di replica sono operativi. Non è richiesta alcuna azione.
- Se un sito preferito non può contattare direttamente il proprio partner o tramite il mediatore, presuppone che il partner sia effettivamente non disponibile oppure è stato isolato e ha portato i percorsi LUN offline. Il sito preferito procede quindi al rilascio dello stato RPO=0 e continua l'elaborazione dell'io in lettura e in scrittura.
- Se un sito non preferito non può contattare direttamente il proprio partner, ma può contattarlo tramite il mediatore, prenderà i suoi percorsi offline e attenderà il ritorno della connessione di replica.
- Se un sito non preferito non può contattare direttamente il proprio partner o tramite un mediatore operativo, supporterà che il partner sia effettivamente non disponibile, oppure che sia stato isolato e che abbia portato i percorsi LUN offline. Il sito non preferito procede quindi al rilascio dello stato RPO=0 e continua l'elaborazione dell'i/o in lettura e scrittura. Assumerà il ruolo dell'origine della replica e diventerà il nuovo sito preferito.

Se il mediatore non è completamente disponibile:

- In caso di guasto dei servizi di replica per qualsiasi motivo, il sito preferito rilascerà lo stato RPO=0 e riprenderà l'elaborazione i/o in lettura e scrittura. Il sito non preferito prenderà i suoi percorsi offline.
- Il guasto del sito preferito causerà un'interruzione poiché il sito non preferito non sarà in grado di verificare che il sito opposto sia effettivamente offline e quindi non sarebbe sicuro per il sito non preferito riprendere i servizi.

Ripristino del servizio

SnapMirror è a riparazione automatica. La sincronizzazione attiva di SnapMirror rileva automaticamente la presenza di una relazione di replica difettosa e la riporta allo stato RPO=0. Una volta ristabilita la replica sincrona, i percorsi torneranno online.

In molti casi, le applicazioni in cluster rilevano automaticamente la restituzione dei percorsi non riusciti e tali applicazioni tornano online. In altri casi, potrebbe essere necessaria una scansione SAN a livello di host oppure potrebbe essere necessario riportare le applicazioni online manualmente.

Dipende dall'applicazione e dal modo in cui è configurata, e in generale tali attività possono essere facilmente automatizzate. La sincronizzazione attiva di SnapMirror stessa risolve automaticamente il problema e non richiede alcun intervento da parte dell'utente per ripristinare l'RPO = 0 operazioni di storage una volta ripristinata l'alimentazione e la connettività.

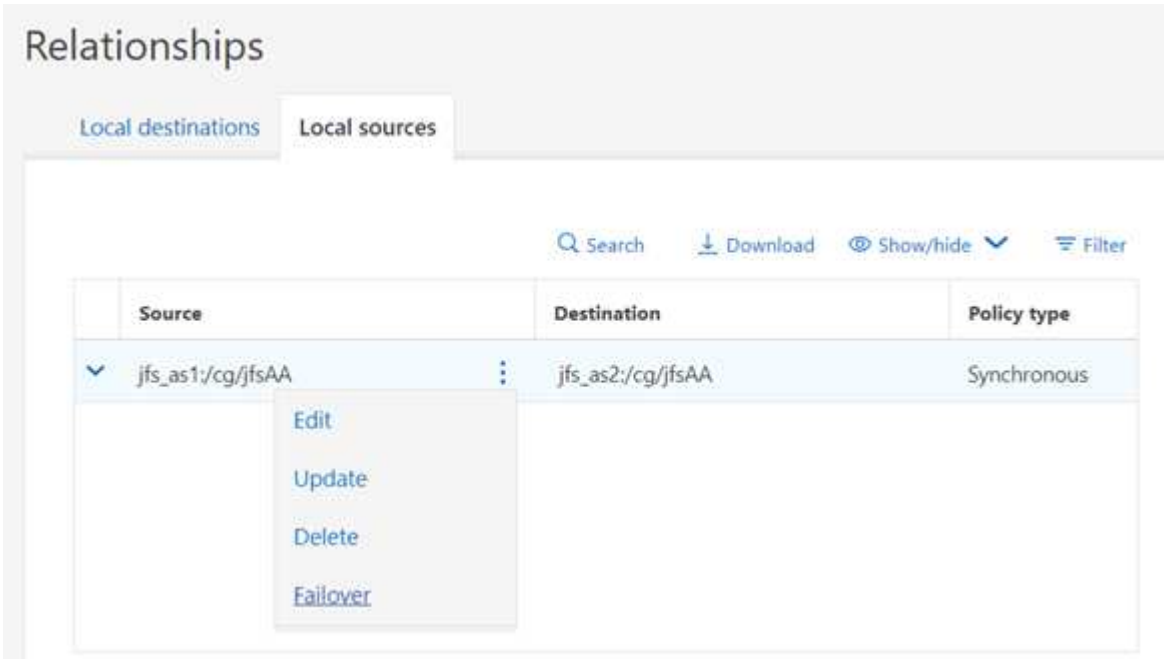
Failover manuale

Il termine "failover" non si riferisce alla direzione della replica con la sincronizzazione attiva di SnapMirror perché si tratta di una tecnologia di replica bidirezionale. "Failover" si riferisce invece al sistema di storage preferito in caso di guasto.

Ad esempio, è possibile eseguire un failover per modificare il sito preferito prima di chiudere un sito per la manutenzione o prima di eseguire un test di DR.

La modifica del sito preferito richiede un'operazione semplice. I/o si fermeranno per un secondo o due come autorità sugli switch del comportamento di replica tra i cluster, ma in caso contrario i/o non vengono influenzati.

Esempio di GUI:



Esempio di modifica tramite l'interfaccia 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
```

Migrazione dei database Oracle

Panoramica

L'utilizzo delle funzionalità di una nuova piattaforma di storage impone un requisito inevitabile e prevede il posizionamento dei dati nel nuovo sistema di storage. ONTAP semplifica il processo di migrazione, inclusi aggiornamenti e migrazioni da ONTAP a ONTAP, importazioni di LUN esterne e procedure per l'utilizzo diretto del sistema operativo host o del software di database Oracle.

Questa documentazione sostituisce il report tecnico precedentemente pubblicato *TR-4534: Migrazione dei database Oracle in sistemi di storage NetApp*

Nel caso di un nuovo progetto di database, questo non rappresenta un problema, poiché gli ambienti di database e applicazioni sono stati costruiti in sede. La migrazione, tuttavia, pone sfide speciali in relazione all'interruzione del business, al tempo necessario per il completamento della migrazione, alle competenze

necessarie e alla minimizzazione del rischio.

Script

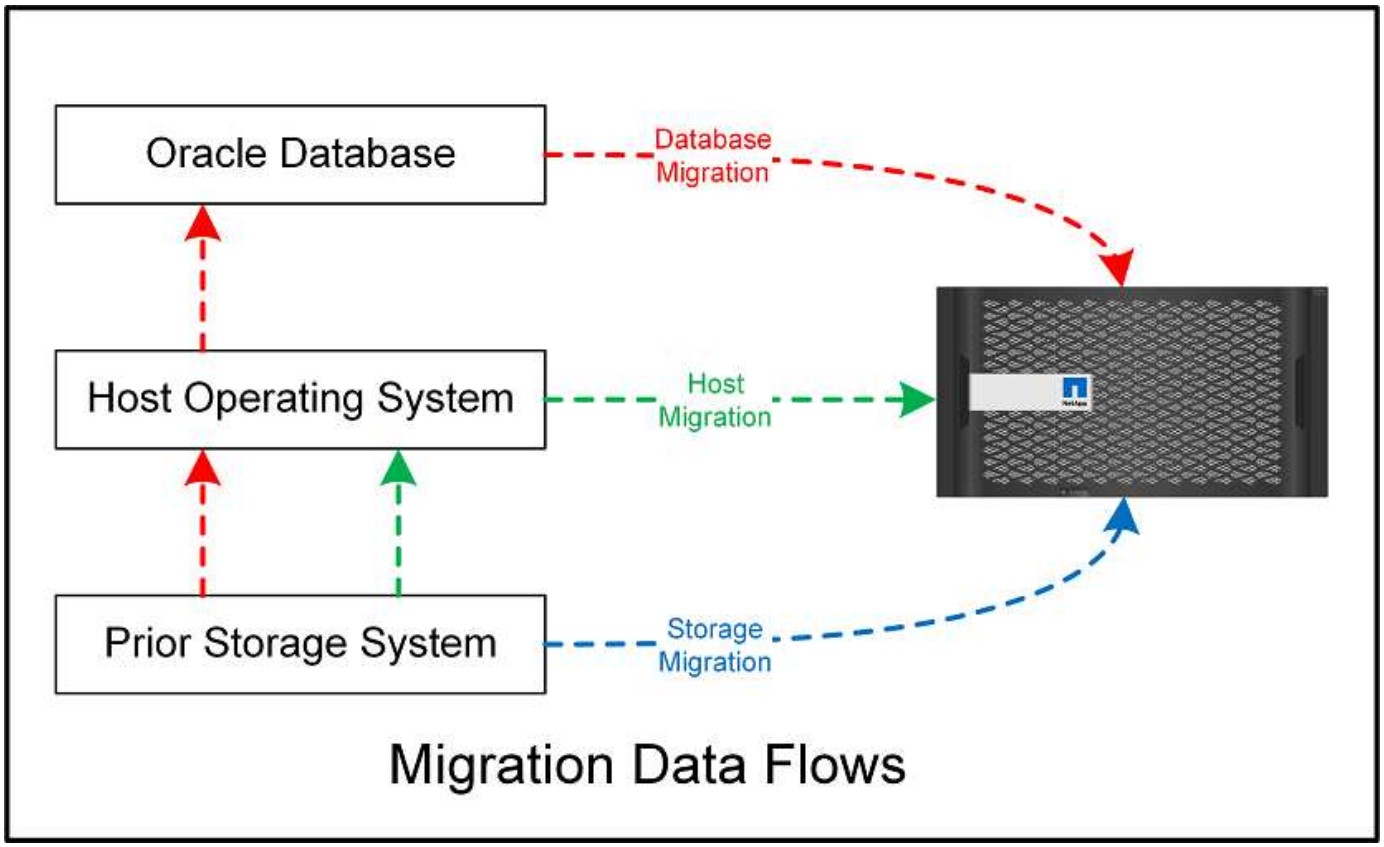
La presente documentazione contiene script di esempio. Questi script forniscono metodi di esempio per automatizzare vari aspetti della migrazione per ridurre la possibilità di errori da parte degli utenti. Gli script possono ridurre le richieste generali del personale IT responsabile della migrazione e accelerare il processo complessivo. Questi script sono ricavati da progetti di migrazione effettivi eseguiti dai servizi di assistenza professionale NetApp e dai partner NetApp. Nella presente documentazione sono riportati alcuni esempi del loro utilizzo.

Pianificazione della migrazione

La migrazione dei dati Oracle può avvenire a uno di tre livelli: Database, host o storage array.

Le differenze risiedono in quale componente della soluzione globale è responsabile dello spostamento dei dati: Il database, il sistema operativo host o il sistema di archiviazione.

La figura riportata di seguito mostra un esempio dei livelli di migrazione e del flusso di dati. In caso di migrazione a livello di database, i dati vengono spostati dal sistema di storage originale ai livelli di host e database nel nuovo ambiente. La migrazione a livello di host è simile, ma i dati non passano attraverso il livello di applicazione e vengono invece scritti nella nuova posizione utilizzando i processi degli host. Infine, con la migrazione a livello di storage, un array come un sistema NetApp FAS si occupa dello spostamento dei dati.



Una migrazione a livello di database si riferisce generalmente all'utilizzo di Oracle log shipping attraverso un database di standby per completare una migrazione a livello di Oracle. Le migrazioni a livello di host vengono eseguite utilizzando le funzionalità native della configurazione del sistema operativo host. Questa configurazione include le operazioni di copia dei file utilizzando comandi quali cp, tar e Oracle Recovery

Manager (RMAN) o un gestore del volume logico (LVM) per spostare i byte sottostanti di un file system. Oracle Automatic Storage Management (ASM) è classificato come capacità a livello di host perché viene eseguito al di sotto del livello dell'applicazione di database. ASM sostituisce il normale volume manager logico su un host. Infine, i dati possono essere migrati a livello di storage array, il che significa sotto il livello del sistema operativo.

Considerazioni sulla pianificazione

La scelta migliore per la migrazione dipende da una combinazione di fattori, inclusa la dimensione dell'ambiente da migrare, la necessità di evitare il downtime e lo sforzo complessivo necessario per eseguire la migrazione. Ovviamente, i database di grandi dimensioni richiedono più tempo e lavoro per la migrazione, ma la complessità di una migrazione di questo tipo è minima. I database di piccole dimensioni possono essere migrati rapidamente, ma se ne devono migrare migliaia, la portata dello sforzo può creare complicazioni. Infine, più grande è il database, più probabile è che l'IT sia business-critical, generando la necessità di ridurre al minimo i downtime mantenendo un percorso di back-out.

Alcune considerazioni per la pianificazione di una strategia di migrazione sono discusse qui.

Dimensioni dei dati

Le dimensioni dei database da migrare influiscono ovviamente sulla pianificazione della migrazione, sebbene le dimensioni non influiscano necessariamente sul tempo di cutover. Quando è necessario migrare una grande quantità di dati, l'aspetto più importante è la larghezza di banda. Le operazioni di copia vengono in genere eseguite con un efficiente i/o sequenziale Come stima conservativa, si presuppone un utilizzo del 50% della larghezza di banda della rete disponibile per le operazioni di copia. Ad esempio, una porta FC da 8GB GB può trasferire in teoria circa 800Mbps GB. Ipotizzando un utilizzo del 50%, è possibile copiare un database a una velocità di circa 400Mbps KB. Pertanto, un database 10TB può essere copiato in circa sette ore a questa velocità.

La migrazione su distanze più lunghe in genere richiede un approccio più creativo, ad esempio il processo di distribuzione dei log illustrato nella "Spostamento file dati online". Le reti IP a lunga distanza raramente dispongono di larghezza di banda in qualsiasi punto vicino alle velocità LAN o SAN. In un caso, NetApp ha assistito alla migrazione a lunga distanza di un database 220TB con tassi di generazione di log di archiviazione molto elevati. L'approccio scelto per il trasferimento dei dati era la spedizione giornaliera dei nastri, perché questo metodo offriva la massima larghezza di banda possibile.

Numero di database

In molti casi, il problema dello spostamento di una grande quantità di dati non è la dimensione dei dati, ma piuttosto la complessità della configurazione che supporta il database. Semplicemente sapere che 50TB database devono essere migrati non è sufficiente. Può essere un singolo database mission-critical 50TB, una raccolta di 4 database legacy 000 o un mix di dati di produzione e non. In alcuni casi, gran parte dei dati è costituita da cloni di un database di origine. Non è necessario migrare questi cloni perché possono essere facilmente ricreati, specialmente quando la nuova architettura è progettata per sfruttare i volumi FlexClone di NetApp.

Per la pianificazione della migrazione, è necessario comprendere il numero dei database interessati e la priorità da assegnare. Con l'aumento del numero di database, l'opzione di migrazione preferita tende a essere più bassa e più bassa nello stack. Ad esempio, la copia di un singolo database può essere eseguita facilmente con RMAN e con una breve interruzione del servizio. Si tratta di una replica a livello di host.

Se ci sono 50 database, potrebbe essere più facile evitare di impostare una nuova struttura del file system per ricevere una copia RMAN e spostare invece i dati sul posto. Questo processo può essere eseguito sfruttando la migrazione LVM basata su host per spostare i dati dalle vecchie LUN ai nuovi LUN. In tal modo, la responsabilità viene trasferita dal team di amministratori del database (DBA) al team del sistema operativo e,

di conseguenza, i dati vengono migrati in modo trasparente rispetto al database. La configurazione del file system rimane invariata.

Infine, se occorre migrare 500 database su 200 server, è possibile utilizzare opzioni basate sullo storage come la funzionalità FLI (ONTAP Foreign LUN Import) per eseguire una migrazione diretta delle LUN.

Requisiti di riarchitettura

In genere, per sfruttare le funzionalità del nuovo storage array è necessario modificare il layout dei file del database; tuttavia, non sempre questo avviene. Ad esempio, le funzionalità degli array all-flash EF-Series sono rivolte principalmente alle performance e all'affidabilità della SAN. Nella maggior parte dei casi, i database possono essere migrati su un array EF-Series senza particolari considerazioni sul layout dei dati. Gli unici requisiti sono IOPS elevati, bassa latenza e solida affidabilità. Sebbene esistano Best practice correlate a fattori quali la configurazione RAID o Dynamic Disk Pools, i progetti EF-Series raramente richiedono modifiche significative all'architettura dello storage generale per sfruttare tali funzionalità.

Al contrario, la migrazione a ONTAP richiede in genere una maggiore considerazione del layout del database per garantire che la configurazione finale fornisca il massimo valore. In sé, ONTAP offre molte funzionalità per un ambiente di database, anche senza interventi specifici sull'architettura. Soprattutto, offre la possibilità di migrare senza interruzioni al nuovo hardware quando l'hardware attuale termina la sua vita utile. In generale, la migrazione a ONTAP è l'ultima migrazione che è necessario eseguire. L'hardware successivo viene aggiornato e i dati vengono migrati senza interruzioni sui nuovi supporti.

Con una certa pianificazione, ancora più benefici sono disponibili. Le considerazioni più importanti riguardano l'uso delle istantanee. Le snapshot sono la base per l'esecuzione di backup, ripristini e operazioni di cloning quasi istantanei. Come esempio della potenza delle istantanee, il più grande utilizzo noto è con un singolo database di 996TB in esecuzione su circa 250 LUN su 6 controller. È possibile eseguire il backup di questo database in 2 minuti, ripristinarlo in 2 minuti e clonarlo in 15 minuti. Tra gli altri benefici, è inclusa la capacità di spostare i dati nel cluster in risposta alle variazioni del carico di lavoro e all'applicazione di controlli di qualità del servizio (QoS) per offrire performance buone e coerenti in un ambiente multi-database.

Tecnologie come controlli della QoS, trasferimento dei dati, snapshot e cloning funzionano praticamente in ogni configurazione. Tuttavia, un certo pensiero è generalmente richiesto per elevare i benefici. In alcuni casi, i layout dello storage del database possono richiedere modifiche di progettazione per massimizzare l'investimento nel nuovo storage array. Tali modifiche di progettazione possono influire sulla strategia di migrazione perché le migrazioni basate su host o su storage replicano il layout dei dati originale. Per completare la migrazione e offrire un layout dei dati ottimizzato per ONTAP potrebbero essere necessari ulteriori passaggi. Le procedure illustrate nella "Panoramica delle procedure di migrazione Oracle" in seguito, dimostrare alcuni metodi non solo per migrare un database, ma anche per eseguirne la migrazione nel layout finale ottimale con il minimo sforzo.

Tempo di cutover

Occorre determinare il disservizio massimo consentito del servizio durante il cutover. È un errore comune presumere che l'intero processo di migrazione causi interruzioni. È possibile eseguire numerose attività prima dell'inizio di qualsiasi interruzione del servizio e completare la migrazione senza interruzioni o black-out attraverso diverse opzioni. Anche quando è inevitabile un'interruzione, è comunque necessario definire il fuori servizio massimo consentito poiché la durata del tempo di cutover varia da procedura a procedura.

Ad esempio, la copia di un database 10TB richiede in genere circa sette ore. Se le esigenze aziendali rendono possibile un'interruzione di sette ore, la copia dei file è un'opzione semplice e sicura per la migrazione. Se cinque ore sono inaccettabili, un semplice log-processo di spedizione (vedere "Log shipping di Oracle") può essere impostato con il minimo sforzo per ridurre il tempo di cutover a circa 15 minuti. Durante questo periodo, un amministratore di database può completare il processo. Se 15 minuti sono inaccettabili, è possibile automatizzare il processo di cutover finale tramite script per ridurre il tempo di cutover a pochi minuti. È

sempre possibile accelerare una migrazione, anche se ciò comporta costi di tempo e lavoro. Gli obiettivi del tempo di cutover devono basarsi su ciò che è accettabile per l'azienda.

Percorso di ritorno

Nessuna migrazione è completamente priva di rischi. Anche se la tecnologia funziona perfettamente, c'è sempre la possibilità di errori da parte dell'utente. Il rischio associato a un percorso di migrazione scelto deve essere preso in considerazione insieme alle conseguenze di una migrazione non riuscita. Ad esempio, la capacità di migrazione trasparente dello storage online di Oracle ASM è una delle sue caratteristiche principali e questo metodo è uno dei più affidabili. Tuttavia, i dati vengono copiati irreversibilmente con questo metodo. Nel caso altamente improbabile in cui si verifichi un problema con ASM, non esiste un facile percorso di back-out. L'unica opzione è ripristinare l'ambiente originale o utilizzare ASM per riportare la migrazione ai LUN originali. Il rischio può essere minimizzato, ma non eliminato, eseguendo un backup di tipo snapshot sul sistema di storage originale, supponendo che il sistema sia in grado di eseguire tale operazione.

Prova

Alcune procedure di migrazione devono essere verificate completamente prima dell'esecuzione. La necessità di migrazione e verifica del processo di cutover è una richiesta comune con i database mission-critical per i quali la migrazione deve avere successo e il downtime deve essere ridotto al minimo. Inoltre, i test di accettazione da parte dell'utente sono spesso inclusi come parte del lavoro di post-migrazione e il sistema complessivo può essere riportato in produzione solo dopo il completamento di questi test.

In caso di necessità di prove, diverse funzionalità di ONTAP possono rendere il processo molto più semplice. In particolare, le istantanee possono ripristinare un ambiente di test e creare rapidamente più copie di un ambiente di database efficienti in termini di spazio.

Procedure

Panoramica

Sono disponibili molte procedure per il database di migrazione Oracle. La giusta dipende dalle vostre esigenze aziendali.

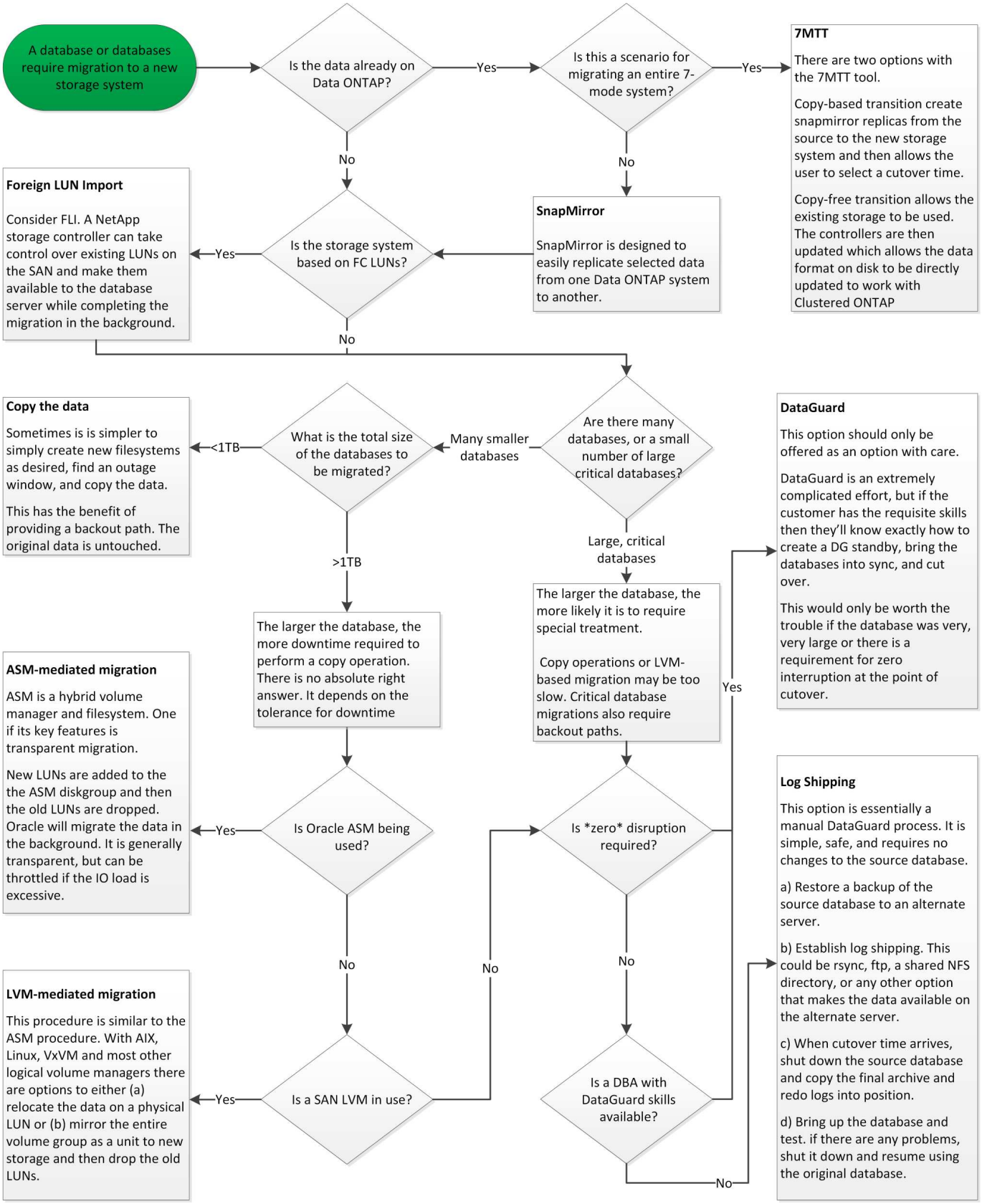
In molti casi, gli amministratori di sistema e i DBA dispongono dei propri metodi preferiti per trasferire i dati dei volumi fisici, eseguire il mirroring e il demirroring o utilizzare Oracle RMAN per copiare i dati.

Queste procedure vengono fornite principalmente come guida per il personale IT meno esperto di alcune delle opzioni disponibili. Inoltre, vengono illustrate le attività, i requisiti di tempo e le richieste di competenze per ogni approccio alla migrazione. Ciò consente ad altre parti, come NetApp e i servizi professionali dei partner o i responsabili dell'IT, di apprezzare più pienamente i requisiti di ogni procedura.

Non esiste un'unica Best practice per la creazione di una strategia di migrazione. La creazione di un piano richiede prima di tutto la comprensione delle opzioni di disponibilità e quindi la selezione del metodo più adatto alle esigenze dell'azienda. La figura seguente illustra le considerazioni di base e le conclusioni tipiche dei clienti, ma non è applicabile a tutte le situazioni.

Ad esempio, un passaggio solleva il problema della dimensione totale del database. Il passaggio successivo dipende dal fatto che il database sia maggiore o minore di 1TB. I passaggi consigliati sono solo questi: Consigli basati su pratiche tipiche del cliente. La maggior parte dei clienti non utilizzerebbe DataGuard per copiare un database di piccole dimensioni, ma alcuni potrebbero farlo. La maggior parte dei clienti non tenterebbe di copiare un database 50TB per il tempo necessario, ma alcuni potrebbero avere una finestra di manutenzione sufficientemente grande da consentire tale operazione.

Il diagramma di flusso riportato di seguito mostra i tipi di considerazioni sul percorso di migrazione più adatto. È possibile fare clic con il pulsante destro del mouse sull'immagine e aprirla in una nuova scheda per migliorare la leggibilità.



.

Spostamento file dati online

Oracle 12cR1 e versioni successive includono la possibilità di spostare un file dati mentre il database rimane online. Inoltre funziona tra diversi tipi di filesystem. Ad esempio, è possibile spostare un file dati da un filesystem xfs ad ASM. Questo metodo non viene generalmente utilizzato su larga scala a causa del numero di operazioni singole di spostamento del file di dati che sarebbero necessarie, ma è un'opzione che vale la pena considerare con database più piccoli con meno file di dati.

Inoltre, il semplice spostamento di un file dati è un'ottima opzione per la migrazione di parti di database esistenti. Ad esempio, è possibile ricollocare i file di dati meno attivi in uno storage più conveniente, ad esempio un volume FabricPool che consente di memorizzare blocchi inattivi in Object Store.

Migrazione a livello di database

La migrazione a livello di database significa consentire il trasferimento dei dati. In particolare, ciò significa spedizione dei log. Tecnologie come RMAN e ASM sono prodotti Oracle, ma, ai fini della migrazione, operano a livello di host in cui copiano i file e gestiscono i volumi.

Spedizione dei log

La base per la migrazione a livello di database è il log di archivio di Oracle, che contiene un registro delle modifiche apportate al database. Nella maggior parte dei casi, un registro di archiviazione fa parte di una strategia di backup e ripristino. Il processo di ripristino inizia con il ripristino di un database e quindi la riproduzione di uno o più log di archivio per portare il database allo stato desiderato. Questa stessa tecnologia di base può essere utilizzata per eseguire una migrazione con interruzioni delle operazioni minime o nulle. Cosa ancora più importante, questa tecnologia consente la migrazione senza intaccare il database originale, preservando un percorso di back-out.

Il processo di migrazione inizia con il ripristino di un backup del database su un server secondario. È possibile farlo in vari modi, ma la maggior parte dei clienti utilizza la normale applicazione di backup per ripristinare i file di dati. Una volta ripristinati i file di dati, gli utenti stabiliscono un metodo per la distribuzione dei log. L'obiettivo è creare un feed costante di log di archivio generati dal database primario e riprodurli sul database ripristinato per mantenerli entrambi vicini allo stesso stato. Quando arriva il tempo di cutover, il database di origine viene completamente arrestato e i log di archivio finali e, in alcuni casi, i log di redo vengono copiati e riprodotti. È fondamentale che i log di ripristino vengano presi in considerazione anche perché potrebbero contenere alcune delle transazioni finali impegnate.

Una volta trasferiti e riprodotti questi log, entrambi i database sono coerenti l'uno con l'altro. A questo punto, la maggior parte dei clienti esegue alcuni test di base. In caso di errori durante il processo di migrazione, la riproduzione del registro dovrebbe segnalare errori e errori. È comunque consigliabile eseguire alcuni test rapidi basati su query note o su attività guidate dalle applicazioni per verificare che la configurazione sia ottimale. È inoltre pratica comune creare una tabella di test finale prima di chiudere il database originale per verificare se è presente nel database migrato. Questa operazione garantisce che non siano stati commessi errori durante la sincronizzazione finale del registro.

Una semplice migrazione log-shipping può essere configurata fuori banda rispetto al database originale, il che lo rende particolarmente utile per i database mission-critical. Non sono richieste modifiche alla configurazione per il database di origine e il ripristino e la configurazione iniziale dell'ambiente di migrazione non hanno alcun effetto sulle operazioni di produzione. Una volta configurato, il log shipping pone alcune richieste di i/o sui server di produzione. Tuttavia, il log shipping è costituito da semplici letture sequenziali dei registri di archivio, che hanno scarse probabilità di influire sulle prestazioni del database di produzione.

La distribuzione dei log si è dimostrata particolarmente utile per progetti di migrazione a lunga distanza e ad alta velocità di cambiamento. In un'istanza, è stata eseguita la migrazione di un singolo database 220TB in una nuova posizione a circa 500 km di distanza. La velocità di modifica era estremamente elevata e le

restrizioni di sicurezza impedivano l'utilizzo di una connessione di rete. La spedizione dei log è stata eseguita utilizzando nastro e corriere. Una copia del database di origine è stata inizialmente ripristinata utilizzando le procedure descritte di seguito. Quindi, i registri sono stati spediti settimanalmente tramite corriere fino al momento del cutover, al momento della consegna del set finale di nastri e dell'applicazione dei registri al database di replica.

Oracle DataGuard

In alcuni casi, è garantito un ambiente DataGuard completo. Non è corretto utilizzare il termine DataGuard per fare riferimento a qualsiasi configurazione del database di standby o di distribuzione dei log. Oracle DataGuard è un framework completo per la gestione della replica dei database, ma non è una tecnologia di replica. Il vantaggio principale di un ambiente DataGuard completo in uno sforzo di migrazione è lo switchover trasparente da un database all'altro. DataGuard consente inoltre uno switchover trasparente nel database originale in caso di problemi, ad esempio problemi di prestazioni o connettività di rete nel nuovo ambiente. Un ambiente DataGuard completamente configurato richiede la configurazione non solo del livello del database ma anche delle applicazioni in modo che le applicazioni siano in grado di rilevare una modifica nella posizione del database primario. In generale, non è necessario utilizzare DataGuard per completare una migrazione, ma alcuni clienti hanno una vasta esperienza DataGuard in-house e già si affidano a essa per le attività di migrazione.

Riarchitettura

Come discusso in precedenza, per sfruttare le funzionalità avanzate degli storage array è talvolta necessario modificare il layout del database. Inoltre, una modifica nel protocollo di storage, come il passaggio da ASM a un file system NFS, altera necessariamente il layout del file system.

Uno dei principali vantaggi dei metodi di distribuzione dei log, incluso DataGuard, è che la destinazione di replica non deve corrispondere all'origine. Non vi sono problemi con l'utilizzo di un approccio di log-shipping per migrare da ASM a un normale file system o viceversa. Il layout preciso dei file di dati può essere modificato a destinazione per ottimizzare l'uso della tecnologia Pluggable Database (PDB) o per impostare i controlli QoS in modo selettivo su determinati file. In altre parole, un processo di migrazione basato sul log shipping consente di ottimizzare il layout dello storage del database in modo semplice e sicuro.

Risorse dei server

Un limite alla migrazione a livello di database è la necessità di un secondo server. Questo secondo server può essere utilizzato in due modi:

1. È possibile utilizzare il secondo server come nuova casa permanente per il database.
2. È possibile utilizzare il secondo server come server di staging temporaneo. Una volta completata e testata la migrazione dei dati nel nuovo storage array, i file system LUN o NFS vengono disconnessi dal server di staging e riconnessi al server originale.

La prima opzione è la più semplice, ma l'utilizzo potrebbe non essere possibile in ambienti molto grandi che richiedono server molto potenti. La seconda opzione richiede ulteriore lavoro per riportare i file system nella posizione originale. Si tratta di una semplice operazione in cui NFS viene utilizzato come protocollo storage, poiché i file system possono essere smontati dal server di staging e rimontati sul server originale.

I file system basati su blocchi richiedono lavoro extra per l'aggiornamento dello zoning FC o degli iSCSI initiator. Con la maggior parte dei gestori di volumi logici (incluso ASM), i LUN vengono automaticamente rilevati e portati online una volta resi disponibili sul server originale. Tuttavia, alcune implementazioni di file system e LVM potrebbero richiedere più lavoro per esportare e importare i dati. La procedura precisa può variare, ma in genere è facile stabilire una procedura semplice e ripetibile per completare la migrazione e ripristinare i dati sul server originale.

Sebbene sia possibile impostare la distribuzione dei log e replicare un database all'interno di un singolo ambiente server, la nuova istanza deve avere un SID di processo diverso per riprodurre i log. È possibile visualizzare temporaneamente il database con un diverso gruppo di ID di processo con un SID diverso e modificarlo in un secondo momento. Tuttavia, questo può portare a numerose e complicate attività di gestione ed espone l'ambiente di database al rischio di errori dell'utente.

Migrazione a livello di host

Migrare i dati a livello di host significa utilizzare il sistema operativo host e le utility associate per completare la migrazione. Questo processo include qualsiasi utility che copia i dati, inclusi Oracle RMAN e Oracle ASM.

Copia dei dati

Il valore di un'operazione di copia semplice non deve essere sottovalutato. Le moderne infrastrutture di rete sono in grado di spostare i dati a velocità misurate in gigabyte al secondo, mentre le operazioni di copia dei file si basano su un efficiente i/o di lettura e scrittura sequenziale L'interruzione è inevitabile con un'operazione di copia dell'host rispetto alla spedizione dei log, ma la migrazione non riguarda solo lo spostamento dei dati. In genere sono incluse le modifiche alla rete, il tempo di riavvio del database e i test post-migrazione.

Il tempo effettivo richiesto per copiare i dati potrebbe non essere significativo. Inoltre, l'operazione di copia preserva un percorso di back-out garantito perché i dati originali non vengono intatti. In caso di problemi durante il processo di migrazione, è possibile riattivare i file system originali con i dati originali.

Riformulazione

Replatforming si riferisce a una modifica del tipo di CPU. Quando un database viene migrato da una piattaforma Solaris, AIX o HP-UX tradizionale a x86 Linux, i dati devono essere riformattati a causa delle modifiche apportate all'architettura della CPU. Le CPU SPARC, IA64 e POWER sono note come grandi processori endian, mentre le architetture x86 e x86_64 sono note come Little endian. Di conseguenza, alcuni dati all'interno dei file di dati Oracle vengono ordinati in modo diverso a seconda del processore in uso.

Tradizionalmente, i clienti utilizzano DataPump per replicare i dati su più piattaforme. DataPump è un'utilità che crea un tipo speciale di esportazione dei dati logici che può essere importata più rapidamente nel database di destinazione. Poiché crea una copia logica dei dati, DataPump lascia alle spalle le dipendenze dell'endianness del processore. Anche se alcuni clienti usano DataPump per il replatform, con Oracle 11g è ora disponibile un'opzione più rapida: Tablespace trasportabili su più piattaforme. Questo avanzamento consente di convertire un tablespace in un diverso formato endian. Si tratta di una trasformazione fisica che offre prestazioni migliori rispetto a un'esportazione DataPump, che deve convertire i byte fisici in dati logici e quindi riconvertirli in byte fisici.

Una discussione completa su DataPump e tablespace trasportabili va oltre la documentazione relativa al NetApp dell'ambito, ma NetApp offre alcuni consigli basati sulla nostra esperienza nell'assistenza ai clienti durante la migrazione a un nuovo log di storage array con una nuova architettura della CPU:

- Se si utilizza DataPump, il tempo necessario per completare la migrazione deve essere misurato in un ambiente di test. A volte i clienti vengono sorpresi del tempo necessario per completare la migrazione. Questo downtime aggiuntivo e inatteso può causare interruzioni delle attività.
- Molti clienti credono erroneamente che gli spazi di tabella trasportabili su più piattaforme non richiedano la conversione dei dati. Quando si utilizza una CPU con un endian diverso, viene utilizzato un RMAN `convert` l'operazione deve essere eseguita sui file di dati in anticipo. Non si tratta di un'operazione istantanea. In alcuni casi, il processo di conversione può essere accelerato avendo più thread che operano su file di dati diversi, ma il processo di conversione non può essere evitato.

Migrazione guidata dal volume logico

Le LVM funzionano prendendo un gruppo di uno o più LUN e suddividendoli in piccole unità generalmente denominate estensioni. Il pool di estensioni viene quindi utilizzato come origine per creare volumi logici essenzialmente virtualizzati. Questo livello di virtualizzazione offre valore in vari modi:

- I volumi logici possono utilizzare estensioni tratte da più LUN. Quando un file system viene creato su un volume logico, può utilizzare le funzionalità con le performance complete di tutte le LUN. Inoltre, promuove il caricamento uniforme di tutte le LUN nel gruppo di volumi, offrendo performance più prevedibili.
- I volumi logici possono essere ridimensionati aggiungendo e, in alcuni casi, rimuovendo le estensioni. Il ridimensionamento di un file system su un volume logico avviene in genere senza interruzione delle attività.
- È possibile migrare i volumi logici senza interruzioni spostando le estensioni sottostanti.

La migrazione tramite LVM funziona in due modi: Spostare un'estensione o specchiare/demirrorizzare un'estensione. La migrazione LVM utilizza l'efficiente i/o sequenziale a blocchi di grandi dimensioni e solo raramente crea problemi di performance. In tal caso, sono solitamente disponibili opzioni per la riduzione della velocità di i/O. In tal modo, si aumenta il tempo necessario per completare la migrazione, riducendo al contempo il carico di i/o sui sistemi host e di storage.

Specchiatura e demirrorazione

Alcuni gestori di volumi, come AIX LVM, consentono all'utente di specificare il numero di copie per ogni estensione e di controllare quali periferiche ospitano ciascuna copia. La migrazione viene eseguita prelevando un volume logico esistente, eseguendo il mirroring delle estensioni sottostanti nei nuovi volumi, attendendo la sincronizzazione delle copie e rilasciando la copia precedente. Se si desidera un percorso di back-out, è possibile creare un'istantanea dei dati originali prima del punto in cui viene rilasciata la copia speculare. In alternativa, è possibile arrestare brevemente il server per mascherare i LUN originali prima di eliminare forzatamente le copie mirror contenute. In tal modo, si preserva una copia recuperabile dei dati nella loro posizione originale.

Estensione della migrazione

Quasi tutti i gestori di volumi consentono la migrazione delle estensioni e talvolta esistono diverse opzioni. Ad esempio, alcuni responsabili di volume consentono a un amministratore di spostare le singole estensioni per un volume logico specifico dal vecchio al nuovo storage. I gestori di volume come Linux LVM2 offrono `pvmove` Che riposiziona tutti gli extent sul dispositivo LUN specificato in un nuovo LUN. Una volta evacuata, la vecchia LUN può essere rimossa.

Il rischio principale per le operazioni è la rimozione delle LUN vecchie e non utilizzate dalla configurazione. È necessario prestare la massima attenzione quando si modifica la suddivisione in zone FC e si rimuovono i dispositivi LUN obsoleti.

Gestione automatica dello storage Oracle

Oracle ASM è un volume manager e un file system logici combinati. A un livello elevato, Oracle ASM prende una raccolta di LUN, le suddivide in piccole unità di allocazione e le presenta come un singolo volume noto come gruppo di dischi ASM. ASM include inoltre la possibilità di eseguire il mirroring del gruppo di dischi impostando il livello di ridondanza. Un volume può essere senza mirror (ridondanza esterna), con mirroring (ridondanza normale) o con mirroring a tre vie (ridondanza elevata). Prestare attenzione durante la configurazione del livello di ridondanza perché non può essere modificato dopo la creazione.

ASM fornisce anche funzionalità di file system. Sebbene il file system non sia visibile direttamente dall'host, il

database Oracle può creare, spostare ed eliminare file e directory in un gruppo di dischi ASM. Inoltre, è possibile navigare nella struttura utilizzando l'utilità asmcmd.

Come per altre implementazioni LVM, Oracle ASM ottimizza le performance di i/o mediante lo striping e il bilanciamento del carico dell'i/o di ciascun file su tutti i LUN disponibili. In secondo luogo, è possibile riposizionare le estensioni sottostanti per consentire sia il ridimensionamento del gruppo di dischi ASM sia la migrazione. Oracle ASM automatizza il processo mediante l'operazione di ribilanciamento. Le nuove LUN vengono aggiunte a un gruppo di dischi ASM e le vecchie LUN vengono eliminate, innescando il trasferimento dell'estensione e la successiva caduta della LUN evacuata dal gruppo di dischi. Questo processo è uno dei metodi di migrazione più comprovati e l'affidabilità di ASM nel fornire una migrazione trasparente è probabilmente la sua caratteristica più importante.

Poiché il livello di mirroring di Oracle ASM è fisso, non può essere utilizzato con il metodo di migrazione mirror e demirroring.

Migrazione a livello di storage

Migrazione a livello di storage: Migrazione al di sotto del livello dell'applicazione e del sistema operativo. In passato, questo a volte significava l'utilizzo di dispositivi specializzati che copiano i LUN a livello di rete, ma queste funzionalità ora si trovano in modo nativo in ONTAP.

SnapMirror

La migrazione di database da un sistema NetApp all'altro viene eseguita quasi universalmente con il software di replica dei dati NetApp SnapMirror. Il processo prevede la configurazione di una relazione di mirroring per i volumi da migrare, in modo che possano essere sincronizzati e quindi in attesa della finestra di cutover. Quando arriva, il database di origine viene arrestato, viene eseguito un aggiornamento finale del mirror e il mirror viene interrotto. I volumi di replica sono quindi pronti per l'uso, montando una directory del file system NFS contenuta oppure rilevando i LUN contenuti e avviando il database.

Il riposizionamento dei volumi in un singolo cluster ONTAP non viene preso in considerazione dalla migrazione, ma piuttosto da una routine `volume move` operazione. SnapMirror viene utilizzato come motore di replica dei dati all'interno del cluster. Questo processo è completamente automatizzato. Non esistono ulteriori passaggi da eseguire per la migrazione quando gli attributi del volume, come la mappatura delle LUN o le autorizzazioni di esportazione NFS, vengono spostati con il volume stesso. Il trasferimento non comporta interruzioni per le operazioni dell'host. In alcuni casi, l'accesso alla rete deve essere aggiornato per garantire che l'accesso ai dati appena ricollocati sia nel modo più efficiente possibile, ma anche queste attività non comportano interruzione delle attività.

Importazione di LUN esterne (FLI)

FLI è una funzione che consente a un sistema Data ONTAP con versione 8,3 o superiore di migrare una LUN esistente da un altro storage array. La procedura è semplice: Il sistema ONTAP viene sottoposto a zoning sull'array di storage esistente come se fosse un qualsiasi altro host SAN. Data ONTAP può quindi controllare le LUN legacy desiderate ed eseguire la migrazione dei dati sottostanti. Inoltre, il processo di importazione utilizza le impostazioni di efficienza del nuovo volume durante la migrazione dei dati, vale a dire che i dati possono essere compressi e deduplicati inline durante il processo di migrazione.

La prima implementazione di FLI in Data ONTAP 8,3 consentiva solo la migrazione offline. Si trattava di un trasferimento molto veloce, ma i dati LUN continuavano a non essere disponibili fino al completamento della migrazione. La migrazione online è stata introdotta in Data ONTAP 8,3.1. Questo tipo di migrazione consente di ridurre al minimo le interruzioni, consentendo a ONTAP di fornire dati LUN durante il processo di trasferimento. Si verifica una breve interruzione mentre l'host viene sottoposto a zoning per l'utilizzo dei LUN tramite ONTAP. Tuttavia, non appena tali modifiche vengono apportate, i dati sono ancora una volta accessibili

e rimangono accessibili per l'intero processo di migrazione.

L'i/o in lettura viene fornito con un proxy tramite ONTAP fino al completamento dell'operazione di copia, mentre l'i/o in scrittura viene scritta in modo sincrono su LUN esterna e ONTAP. Le due copie LUN vengono mantenute sincronizzate in questo modo fino a quando l'amministratore non esegue un cutover completo che rilascia la LUN esterna e non replica più le scritture.

FLI è progettato per funzionare con FC, ma se si desidera passare a iSCSI, la LUN migrata può essere facilmente rimappata come una LUN iSCSI al termine della migrazione.

Tra le caratteristiche di FLI vi è il rilevamento e la regolazione automatici dell'allineamento. In questo contesto, il termine allineamento si riferisce a una partizione su un dispositivo LUN. Per ottenere prestazioni ottimali è necessario allineare l'i/o ai blocchi da 4K KB. Se una partizione viene posizionata su un offset che non è multiplo di 4K, le prestazioni ne risentono.

Esiste un secondo aspetto dell'allineamento che non può essere corretto regolando un offset di partizione, ovvero la dimensione del blocco del file system. Ad esempio, un file system ZFS generalmente utilizza per impostazione predefinita una dimensione di blocco interna di 512 byte. Altri clienti che utilizzano AIX hanno occasionalmente creato file system JFS2 con dimensioni blocco di 512 o 1, 024 byte. Anche se il file system potrebbe essere allineato a un limite di 4K, i file creati all'interno di tale file system non lo sono e le prestazioni ne risentono.

FLI non deve essere usato in queste circostanze. Anche se i dati sono accessibili dopo la migrazione, il risultato sono file system con gravi limitazioni delle prestazioni. In linea di principio, qualsiasi file system che supporti un carico di lavoro di sovrascrittura casuale su ONTAP dovrebbe utilizzare una dimensione del blocco di 4K KB. Ciò è applicabile principalmente a workload come file di dati di database e implementazioni di VDI. La dimensione del blocco può essere identificata utilizzando i comandi del sistema operativo host pertinente.

Ad esempio, su AIX, la dimensione del blocco può essere visualizzata con `lsfs -q`. Con Linux, `xfs_info` e. `tune2fs` può essere utilizzato per `xfs` e. `ext3/ext4`, rispettivamente. Con `zfs`, il comando è `zdb -C`.

Il parametro che controlla la dimensione del blocco è `ashift` e generalmente il valore predefinito è 9, che significa 2^9, o 512 byte. Per prestazioni ottimali, la `ashift` Il valore deve essere 12 (2^12=4K). Questo valore viene impostato al momento della creazione di zpool e non può essere modificato, il che significa che i data zpool con un `ashift` oltre a 12 deve essere eseguita la migrazione copiando i dati in uno zpool appena creato.

Oracle ASM non ha dimensioni dei blocchi fondamentali. L'unico requisito è che la partizione su cui è stato creato il disco ASM sia allineata correttamente.

7-Mode Transition Tool

7-Mode Transition Tool (7MTT) è un'utility di automazione utilizzata per migrare configurazioni 7- Mode di grandi dimensioni a ONTAP. La maggior parte dei clienti che gestiscono i database trovano altri metodi più semplici, in parte perché eseguono di solito la migrazione dei database piuttosto che trasferire l'intero footprint dello storage. Inoltre, i database sono spesso solo una parte di un ambiente di storage più ampio. Pertanto, spesso i database vengono migrati singolarmente, quindi l'ambiente rimanente può essere spostato con 7MTT.

Alcuni clienti con sistemi di storage dedicati a ambienti di database complicati hanno un numero limitato ma significativo di essi. Questi ambienti potrebbero contenere molti volumi, snapshot e numerosi dettagli di configurazione, come autorizzazioni di esportazione, gruppi iniziatori LUN, autorizzazioni utente e configurazione del protocollo Lightweight Directory Access Protocol. In questi casi, le capacità di automazione di 7MTT possono semplificare una migrazione.

7MTT può funzionare in una delle due modalità seguenti:

- **Copy- Based Transition (CBT).** 7MTT con CBT imposta i volumi SnapMirror da un sistema 7- Mode esistente nel nuovo ambiente. Una volta sincronizzati i dati, 7MTT orchestra il processo di cutover.
- **Copy- Free Transition (CFT).** 7MTT con CFT si basa sulla conversione in-place degli shelf di dischi 7-Mode esistenti. I dati non vengono copiati e gli shelf di dischi esistenti possono essere riutilizzati. La configurazione esistente di data Protection ed efficienza dello storage viene preservata.

La differenza principale tra queste due opzioni consiste nel fatto che la transizione senza copie è un approccio a big-bang, in cui tutti gli shelf di dischi collegati alla coppia ha 7- Mode originale devono essere ricollocati nel nuovo ambiente. Non esiste alcuna opzione per spostare un sottoinsieme di shelf. L'approccio basato sulla copia consente lo spostamento dei volumi selezionati. Esiste anche potenzialmente una finestra di cutover più lunga con transizione priva di copie a causa del legame necessario per la riselezione degli shelf di dischi e la conversione dei metadati. In base all'esperienza sul campo, NetApp consiglia di lasciare trascorrere 1 ora per il riposizionamento e il ripristino degli shelf di dischi e tra 15 minuti e 2 ore per la conversione dei metadati.

Migrazione di file dati

È possibile spostare singoli file di dati Oracle con un singolo comando.

Ad esempio, il comando seguente sposta il file dati IOPST.dbf dal filesystem `/oradata2` al filesystem `/oradata3`.

```
SQL> alter database move datafile  '/oradata2/NTAP/IOPS002.dbf' to
'/oradata3/NTAP/IOPS002.dbf';
Database altered.
```

Lo spostamento di un file dati con questo metodo può essere lento, ma in genere non dovrebbe produrre i/o sufficienti da interferire con i carichi di lavoro del database quotidiani. Al contrario, la migrazione tramite il ribilanciamento di ASM può essere eseguita molto più rapidamente, ma con il rischio di rallentare il database globale durante lo spostamento dei dati.

È possibile misurare facilmente il tempo necessario per spostare i file di dati creando un file di dati di test e spostandolo. Il tempo trascorso per l'operazione viene registrato nei dati di v$session:

```
SQL> set linesize 300;
SQL> select elapsed_seconds||':'||message from v$session_longops;
ELAPSED_SECONDS||':'||MESSAGE
------------------------------------------------------------------------
---------------
351:Online data file move: data file 8: 22548578304 out of 22548578304
bytes done
SQL> select bytes / 1024 / 1024 /1024 as GB from dba_data_files where
FILE_ID = 8;
        GB
----------
        21
```

In questo esempio, il file spostato era datafile 8, della dimensione di 21GB GB e della durata di 6 minuti per la migrazione. Il tempo necessario dipende ovviamente dalle funzionalità del sistema di storage, della rete di

storage e dall'attività complessiva del database che si verifica al momento della migrazione.

Spedizione dei log

L'obiettivo di una migrazione utilizzando la distribuzione dei log è creare una copia dei file di dati originali in una nuova posizione e quindi stabilire un metodo per la distribuzione delle modifiche nel nuovo ambiente.

Una volta stabiliti, è possibile automatizzare la spedizione e la riproduzione dei log per mantenere il database di replica ampiamente sincronizzato con l'origine. Ad esempio, un job cron può essere programmato per (a) copiare i log più recenti nella nuova posizione e (b) riprodurli ogni 15 minuti. In questo modo si riduce al minimo l'interruzione al momento del cutover, in quanto è necessario riprodurre non più di 15 minuti dei registri di archivio.

La procedura illustrata di seguito è essenzialmente un'operazione di clonazione del database. La logica illustrata è simile al motore all'interno di NetApp SnapManager per Oracle (SMO) e al plug-in NetApp SnapCenter per Oracle. Alcuni clienti utilizzano la procedura indicata negli script o nei workflow Wfa per le operazioni di cloning personalizzate. Sebbene questa procedura sia più manuale che non utilizzi SMO o SnapCenter, viene comunque rapidamente script e le API di gestione dei dati all'interno di ONTAP semplificano ulteriormente il processo.

Log shipping - dal file system al file system

In questo esempio viene illustrata la migrazione di un database denominato WAFFLE da un normale file system a un altro normale file system situato su un server diverso. Illustra anche l'utilizzo di SnapMirror per eseguire una copia rapida dei file di dati, ma questa non è parte integrante della procedura generale.

Creare il backup del database

Il primo passo consiste nel creare un backup del database. In particolare, questa procedura richiede una serie di file di dati che possono essere utilizzati per la riproduzione del log di archivio.

Ambiente

In questo esempio, il database di origine si trova su un sistema ONTAP. Il metodo più semplice per creare un backup di un database consiste nell'utilizzare uno snapshot. Il database viene messo in modalità di backup a caldo per alcuni secondi mentre un `snapshot create` l'operazione viene eseguita sul volume che ospita i file di dati.

```
SQL> alter database begin backup;
Database altered.
```

```
Cluster01::*> snapshot create -vserver vserver1 -volume jfsc1_oradata
hotbackup
Cluster01::*>
```

```
SQL> alter database end backup;
Database altered.
```

Il risultato è un'istantanea sul disco chiamata `hotbackup` che contiene un'immagine dei file di dati in modalità di backup a caldo. Se combinati con i log di archivio appropriati per rendere i file di dati coerenti, i dati di questa snapshot possono essere utilizzati come base di un ripristino o di un clone. In questo caso, viene replicato sul nuovo server.

Ripristino in un nuovo ambiente

Ora il backup deve essere ripristinato nel nuovo ambiente. Questa operazione può essere eseguita in vari modi, tra cui Oracle RMAN, ripristino da un'applicazione di backup come NetBackup o semplice operazione di copia dei file di dati inseriti in modalità hot backup.

In questo esempio, SnapMirror viene utilizzato per replicare l'hot backup dello snapshot in una nuova posizione.

1. Creare un nuovo volume per ricevere i dati dello snapshot. Inizializzare il mirroring da `jfsc1_oradata` a. `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. Una volta impostato lo stato da SnapMirror, a indicare che la sincronizzazione è completa, aggiornare il mirror in base allo snapshot desiderato,

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields state
source-path             destination-path        state
----------------------- ----------------------- ------------
vserver1:jfsc1_oradata vserver1:vol_oradata SnapMirrored
```

```
Cluster01::*> snapmirror update -destination-path vserver1:vol_oradata
-source-snapshot hotbackup
Operation is queued: snapmirror update of destination
"vserver1:vol_oradata".
```

3. La sincronizzazione può essere verificata visualizzando `newest-snapshot` sul volume speculare.

```
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. Lo specchio può quindi essere rotto.

```
Cluster01::> snapmirror break -destination-path vserver1:vol_oradata
Operation succeeded: snapmirror break for destination
"vserver1:vol_oradata".
Cluster01::>
```

5. Montare il nuovo file system.con i file system basati su blocchi, le procedure precise variano in base al LVM in uso. È necessario configurare lo zoning FC o le connessioni iSCSI. Dopo aver stabilito la connettività ai LUN, comandi come Linux `pvscan` Potrebbe essere necessario per rilevare quali gruppi di volumi o LUN devono essere configurati correttamente per essere rilevati da ASM.

   In questo esempio viene utilizzato un semplice file system NFS. Questo file system può essere montato direttamente.

```
fas8060-nfs1:/vol_oradata        19922944   1639360   18283584   9%
/oradata
fas8060-nfs1:/vol_logs            9961472       128    9961344   1%
/logs
```

Creare un modello di creazione controlfile

Successivamente, è necessario creare un modello controlfile. Il `backup controlfile to trace` comando crea comandi di testo per ricreare un controlfile. In alcuni casi, questa funzione può risultare utile per ripristinare un database da un backup e viene spesso utilizzata con script che eseguono attività come la clonazione dei database.

1. L'output del comando seguente viene utilizzato per ricreare i file di controllo per il database migrato.

```
SQL> alter database backup controlfile to trace as '/tmp/waffle.ctrl';
Database altered.
```

2. Una volta creati i file di controllo, copiarli nel nuovo server.

```
[oracle@jfsc3 tmp]$ scp oracle@jfsc1:/tmp/waffle.ctrl /tmp/
oracle@jfsc1's password:
waffle.ctrl                                               100% 5199
5.1KB/s   00:00
```

File dei parametri di backup

Nel nuovo ambiente è necessario anche un file di parametri. Il metodo più semplice consiste nel creare un pfile dal file spfile o pfile corrente. In questo esempio, il database di origine utilizza un spfile.

```
SQL> create pfile='/tmp/waffle.tmp.pfile' from spfile;
File created.
```

Crea voce oratab

La creazione di una voce oratab è necessaria per il corretto funzionamento di utility come oraenv. Per creare una voce oratab, completare il passaggio seguente.

```
WAFFLE:/orabin/product/12.1.0/dbhome_1:N
```

Preparare la struttura delle directory

Se le directory richieste non sono già presenti, è necessario crearle oppure la procedura di avvio del database non riesce. Per preparare la struttura di directory, completare i seguenti requisiti minimi.

```
[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
```

Aggiornamenti del file dei parametri

1. Per copiare il file dei parametri nel nuovo server, eseguire i seguenti comandi. La posizione predefinita è `$ORACLE_HOME/dbs` directory. In questo caso, il pfile può essere posizionato ovunque. Viene utilizzata solo come fase intermedia del processo di migrazione.

```
[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. Modificare il file come richiesto. Ad esempio, se la posizione del log di archivio è stata modificata, il file pfile deve essere modificato per riflettere la nuova posizione. In questo esempio, vengono ricollocati solo i file di controllo, in parte per distribuirli tra i file system di log e di dati.

```
[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. Al termine delle modifiche, creare un file spfile basato su questo file pfile.

```
SQL> create spfile from pfile='waffle.tmp.pfile';
File created.
```

Ricreare i file di controllo

In una fase precedente, l'output di `backup controlfile to trace` è stato copiato nel nuovo server. La parte specifica dell'uscita richiesta è la `controlfile recreation` comando. Queste informazioni si trovano nel file sotto la sezione contrassegnata `Set #1. NORESETLOGS`. Inizia con la linea `create controlfile reuse database` e dovrebbe includere la parola `noresetlogs`. Termina con il punto e virgola (;).

1. In questa procedura di esempio, il file viene letto come segue.

```
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. Modificare lo script come desiderato per riflettere la nuova posizione dei vari file. Ad esempio, alcuni file di dati noti per supportare un i/o elevato potrebbero essere reindirizzati a un file system su un Tier di storage dalle performance elevate. In altri casi, le modifiche possono essere apportate solo per motivi di amministrazione, ad esempio isolando i file di dati di un PDB in volumi dedicati.
3. In questo esempio, il `DATAFILE` stanza viene lasciata invariata, ma i log di redo vengono spostati in una nuova posizione in `/redo` piuttosto che condividere lo spazio con i log di archivio `/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>
```

Se i file sono posizionati in modo errato o i parametri non sono configurati correttamente, vengono generati errori che indicano ciò che deve essere corretto. Il database è montato, ma non è ancora aperto e non può essere aperto perché i file di dati in uso sono ancora contrassegnati come in modalità di backup a caldo. Per rendere il database coerente, è necessario applicare prima i registri di archiviazione.

Replica iniziale del registro

Per rendere coerenti i file di dati è necessaria almeno un'operazione di risposta del registro. Sono disponibili molte opzioni per la riproduzione dei registri. In alcuni casi, la posizione originale del log di archivio sul server originale può essere condivisa tramite NFS e la risposta del log può essere effettuata direttamente. In altri casi, è necessario copiare i registri di archivio.

Ad esempio, un semplice `scp` l'operazione può copiare tutti i log correnti dal server di origine al server di migrazione:

```
[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
```

Riproduzione del registro iniziale

Una volta che i file si trovano nella posizione del log di archivio, è possibile riprodurli inviando il comando `recover database until cancel` seguito dalla risposta `AUTO` per riprodurre automaticamente tutti i registri disponibili.

```
SQL> recover database until cancel;
ORA-00279: change 382713 generated at 05/24/2016 09:00:54 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_23_912662036.dbf
ORA-00280: change 382713 for thread 1 is in sequence #23
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 405712 generated at 05/24/2016 15:01:05 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_24_912662036.dbf
ORA-00280: change 405712 for thread 1 is in sequence #24
ORA-00278: log file '/logs/WAFFLE/arch/1_23_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
ORA-00278: log file '/logs/WAFFLE/arch/1_30_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_31_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

La risposta finale del log di archivio riporta un errore, ma questo è normale. Il registro indica che `sqlplus` stava cercando un particolare file di registro e non lo ha trovato. Il motivo è, molto probabilmente, che il file di registro non esiste ancora.

Se il database di origine può essere arrestato prima di copiare i registri di archivio, questa operazione deve essere eseguita una sola volta. I log di archivio vengono copiati e riprodotti, quindi il processo può continuare direttamente con il processo di cutover che replica i log di ripristino critici.

Replica e riproduzione incrementale dei log

Nella maggior parte dei casi, la migrazione non viene eseguita immediatamente. Il completamento del processo di migrazione potrebbe richiedere alcuni giorni o addirittura settimane, pertanto i log devono essere inviati continuamente al database di replica e riprodotti. Pertanto, quando arriva il cutover, occorre trasferire e riprodurre minimi dati.

In questo modo è possibile eseguire script in molti modi diversi, ma uno dei metodi più diffusi è l'utilizzo di rsync, un'utilità comune di replica dei file. Il modo più sicuro per usare questa utility è configurarla come demone. Ad esempio, il `rsyncd.conf` file che segue mostra come creare una risorsa chiamata `waffle.arch` A cui si accede con le credenziali utente Oracle e a cui è mappato `/logs/WAFFLE/arch`. Soprattutto, la risorsa è impostata su sola lettura, consentendo la lettura dei dati di produzione, ma non l'alterazione.

```
[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
```

Il seguente comando sincronizza la destinazione del log di archivio del nuovo server con la risorsa rsync `waffle.arch` sul server originale. Il `t` argomento in `rsync - potg` fa sì che l'elenco di file venga confrontato in base alla data e all'ora e che vengano copiati solo i nuovi file. Questo processo fornisce un aggiornamento incrementale del nuovo server. Questo comando può anche essere programmato in cron per essere eseguito regolarmente.

```
[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
```

Una volta ricevuti i registri, è necessario riprodurli. Gli esempi precedenti mostrano l'uso di sqlplus per l'esecuzione manuale `recover database until cancel`, un processo che può essere facilmente automatizzato. Nell'esempio illustrato viene utilizzato lo script descritto nella "Riproduci i registri sul database". Gli script accettano un argomento che specifica il database che richiede un'operazione di riproduzione. Ciò consente di utilizzare lo stesso script in una migrazione di più database.

```
[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
```

Cutover

Quando si è pronti per il passaggio al nuovo ambiente, è necessario eseguire una sincronizzazione finale che includa sia i registri di archivio che i registri di ripristino. Se la posizione originale del log di ripristino non è già nota, è possibile identificarla come segue:

```
SQL> select member from v$logfile;
MEMBER
--------------------------------------------------------------------------------
------
/logs/WAFFLE/redo/redo01.log
/logs/WAFFLE/redo/redo02.log
/logs/WAFFLE/redo/redo03.log
```

1. Arrestare il database di origine.
2. Eseguire una sincronizzazione finale dei registri di archivio sul nuovo server con il metodo desiderato.
3. I log di ripristino di origine devono essere copiati nel nuovo server. In questo esempio, i log di ripristino sono stati spostati in una nuova directory all'indirizzo `/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 questa fase, il nuovo ambiente di database contiene tutti i file necessari per portarlo nello stesso stato dell'origine. I registri di archivio devono essere riprodotti una volta finale.

```
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. Al termine, i log di ripristino devono essere riprodotti. Se il messaggio `Media recovery complete` viene restituito, il processo è riuscito e i database sono sincronizzati e possono essere aperti.

```
SQL> recover database;
Media recovery complete.
SQL> alter database open;
Database altered.
```

Log shipping - da ASM a file system

In questo esempio viene illustrato l'utilizzo di Oracle RMAN per la migrazione di un database. È molto simile all'esempio precedente di distribuzione del log del file system, ma i file su ASM non sono visibili all'host. Le uniche opzioni per la migrazione dei dati presenti sui dispositivi ASM sono il riposizionamento del LUN ASM o l'utilizzo di Oracle RMAN per eseguire le operazioni di copia.

Sebbene RMAN sia un requisito per la copia dei file da Oracle ASM, l'utilizzo di RMAN non è limitato a ASM. RMAN può essere utilizzato per migrare da qualsiasi tipo di storage a qualsiasi altro tipo.

Questo esempio mostra il trasferimento di un database chiamato PANCAKE dallo storage ASM a un file system normale situato su un server diverso nei percorsi `/oradata` e. `/logs`.

Creare il backup del database

Il primo passo consiste nel creare un backup del database da migrare su un server alternativo. Poiché l'origine utilizza Oracle ASM, è necessario utilizzare RMAN. Un semplice backup RMAN può essere eseguito come segue. Questo metodo crea un backup con tag che può essere facilmente identificato da RMAN più avanti nella procedura.

Il primo comando definisce il tipo di destinazione per il backup e la posizione da utilizzare. Il secondo avvia il

backup dei soli file di dati.

```
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
```

Backup controlfile

Un controlfile di backup è necessario più avanti nella procedura per `duplicate database` operazione.

```
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
```

File dei parametri di backup

Nel nuovo ambiente è necessario anche un file di parametri. Il metodo più semplice consiste nel creare un pfile dal file spfile o pfile corrente. In questo esempio, il database di origine utilizza un spfile.

```
RMAN> create pfile='/rman/pancake/pfile' from spfile;
Statement processed
```

Script di ridenominazione file ASM

Diverse posizioni dei file attualmente definite nei file di controllo cambiano quando il database viene spostato. Lo script seguente crea uno script RMAN per semplificare il processo. Questo esempio mostra un database con un numero molto ridotto di file di dati, ma in genere i database contengono centinaia o addirittura migliaia di file di dati.

Questo script si trova in "Conversione da ASM a nome file system" e fa due cose.

In primo luogo, viene creato un parametro per ridefinire le posizioni del log di ripristino chiamate `log_file_name_convert`. Si tratta essenzialmente di un elenco di campi alternati. Il primo campo rappresenta la posizione di un registro di ripristino corrente, mentre il secondo campo rappresenta la posizione sul nuovo server. Il modello viene quindi ripetuto.

La seconda funzione consiste nel fornire un modello per la ridenominazione dei file di dati. Lo script esegue il ciclo dei file di dati, estrae le informazioni sul nome e sul numero del file e lo formatta come uno script RMAN. Quindi fa lo stesso con i file temporanei. Il risultato è un semplice script rman che può essere modificato come desiderato per assicurarsi che i file vengano ripristinati nella posizione desiderata.

```
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.
```

Acquisire l'output di questa schermata. Il `log_file_name_convert` il parametro viene inserito nel file pfile come descritto di seguito. Il file di dati RMAN rinominato e lo script duplicato devono essere modificati di conseguenza per posizionare i file di dati nelle posizioni desiderate. In questo esempio, sono tutti inseriti `/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';
}
```

Preparare la struttura delle directory

Gli script sono quasi pronti per l'esecuzione, ma prima la struttura di directory deve essere in posizione. Se le directory richieste non sono già presenti, è necessario crearle oppure la procedura di avvio del database non riesce. L'esempio riportato di seguito riflette i requisiti minimi.

```
[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
```

Crea voce oratab

Il seguente comando è necessario per il corretto funzionamento di utility come oraenv.

```
PANCAKE:/orabin/product/12.1.0/dbhome_1:N
```

Aggiornamenti dei parametri

Il file pfile salvato deve essere aggiornato per riflettere eventuali modifiche di percorso sul nuovo server. Le modifiche al percorso del file di dati vengono modificate dallo script di duplicazione RMAN e quasi tutti i database richiedono modifiche al `control_files` e. `log_archive_dest` parametri. Potrebbero inoltre essere presenti posizioni dei file di controllo che devono essere modificate e parametri quali `db_create_file_dest` Potrebbe non essere rilevante al di fuori di ASM. Prima di procedere, un DBA esperto deve esaminare attentamente le modifiche proposte.

In questo esempio, le modifiche principali sono le posizioni controlfile, la destinazione di archivio del registro e l'aggiunta di `log_file_name_convert` parametro.

```
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'
```

Dopo la conferma dei nuovi parametri, i parametri devono essere applicati. Esistono diverse opzioni, ma la maggior parte dei clienti crea un file spfile basato sul file pfile di testo.

```
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.
```

Nomount di avvio

Il passaggio finale prima della replica del database consiste nel visualizzare i processi del database ma non nel montare i file. In questa fase, potrebbero manifestarsi problemi con spfile. Se il `startup nomount` comando non riesce a causa di un errore di parametro, è semplice chiudere, correggere il modello pfile, ricaricarlo come spfile, e riprovare.

```
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
```

Duplicare il database

Il ripristino del backup RMAN precedente nella nuova posizione richiede più tempo rispetto ad altre fasi di questo processo. Il database deve essere duplicato senza modificare l'ID del database (DBID) o reimpostare i registri. Ciò impedisce l'applicazione dei registri, operazione necessaria per la sincronizzazione completa delle copie.

Connettersi al database con RMAN come aux ed eseguire il comando duplicato del database utilizzando lo script creato in un passaggio precedente.

```
[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
```

Replica iniziale del registro

A questo punto è necessario inviare le modifiche dal database di origine a una nuova posizione. In tal caso, potrebbe essere necessario eseguire una combinazione di operazioni. Il metodo più semplice sarebbe fare in modo che RMAN nel database di origine scriva i log di archivio in una connessione di rete condivisa. Se una posizione condivisa non è disponibile, un metodo alternativo consiste nell'utilizzare RMAN per scrivere su un file system locale e quindi utilizzare rcp o rsync per copiare i file.

In questo esempio, il `/rman` Directory è una condivisione NFS disponibile sia per il database originale che per quello migrato.

Una questione importante in questo caso è la `disk format` clausola. Il formato del disco del backup è `%h_%e_%a.dbf`, Che significa che è necessario utilizzare il formato del numero di thread, il numero di sequenza e l'ID di attivazione per il database. Anche se le lettere sono diverse, questa corrisponde alla `log_archive_format='%t_%s_%r.dbf` parametro nel pfile. Questo parametro specifica inoltre i log di archivio nel formato di numero di thread, numero di sequenza e ID di attivazione. Il risultato finale è che i backup del file di registro sull'origine utilizzano una convenzione di denominazione prevista dal database. In questo modo, vengono eseguite operazioni come `recover database` molto più semplice perché sqlplus anticipa correttamente i nomi dei log di archivio da riprodurre.

```
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
```

Riproduzione del registro iniziale

Una volta che i file si trovano nella posizione del log di archivio, è possibile riprodurli inviando il comando `recover database until cancel` seguito dalla risposta `AUTO` per riprodurre automaticamente tutti i registri disponibili. Il file dei parametri sta attualmente indirizzando i log di archivio a. `/logs/archive`, Ma non corrisponde alla posizione in cui RMAN è stato utilizzato per salvare i registri. La posizione può essere reindirizzata temporaneamente come segue prima di ripristinare il database.

```
SQL> alter system set log_archive_dest_1='LOCATION=/rman/pancake/logship'
scope=memory;
System altered.
SQL> recover standby database until cancel;
ORA-00279: change 560224 generated at 05/24/2016 03:25:53 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_49_912576125.dbf
ORA-00280: change 560224 for thread 1 is in sequence #49
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 560353 generated at 05/24/2016 03:29:17 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_50_912576125.dbf
ORA-00280: change 560353 for thread 1 is in sequence #50
ORA-00278: log file '/rman/pancake/logship/1_49_912576125.dbf' no longer
needed
for this recovery
...
ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
ORA-00278: log file '/rman/pancake/logship/1_53_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_54_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

La risposta finale del log di archivio riporta un errore, ma questo è normale. L'errore indica che sqlplus stava cercando un particolare file di registro e non lo ha trovato. Il motivo è molto probabile che il file di registro non esista ancora.

Se il database di origine può essere arrestato prima di copiare i registri di archivio, questa operazione deve essere eseguita una sola volta. I log di archivio vengono copiati e riprodotti, quindi il processo può continuare direttamente con il processo di cutover che replica i log di ripristino critici.

Replica e riproduzione incrementale dei log

Nella maggior parte dei casi, la migrazione non viene eseguita immediatamente. Il completamento del processo di migrazione potrebbe richiedere alcuni giorni o addirittura settimane, pertanto i log devono essere inviati continuamente al database di replica e riprodotti. In questo modo si assicura che i dati minimi debbano essere trasferiti e riprodotti all'arrivo del cutover.

Questo processo può essere facilmente gestito tramite script. Ad esempio, è possibile pianificare il seguente comando nel database originale per assicurarsi che la posizione utilizzata per la spedizione dei log venga

aggiornata continuamente.

```
[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.
```

Una volta ricevuti i registri, è necessario riprodurli. Gli esempi precedenti hanno mostrato l'uso di sqlplus per l'esecuzione manuale `recover database until cancel`, che può essere facilmente automatizzato. Nell'esempio illustrato viene utilizzato lo script descritto nella "Replay Logs on Standby Database". Lo script accetta un argomento che specifica il database che richiede un'operazione di riproduzione. Questo processo consente di utilizzare lo stesso script in una migrazione di più database.

```
[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
```

Cutover

Quando si è pronti a passare al nuovo ambiente, è necessario eseguire una sincronizzazione finale. Quando si lavora con i normali file system, è facile assicurarsi che il database migrato sia sincronizzato al 100% rispetto all'originale, poiché i log di ripristino originali vengono copiati e riprodotti. Con ASM non esiste un buon modo per farlo. È possibile recuperare facilmente solo i registri di archivio. Per assicurarsi che i dati non vadano persi, è necessario eseguire con attenzione l'arresto finale del database originale.

1. In primo luogo, la base di dati deve essere chiusa, garantendo che non vengano apportate modifiche. Questa chiusura potrebbe includere la disattivazione delle operazioni pianificate, la chiusura dei listener e/o la chiusura delle applicazioni.
2. Una volta eseguita questa operazione, la maggior parte dei DBA crea una tabella fittizia da utilizzare come indicatore dell'arresto.
3. Forzare l'archiviazione di un registro per assicurarsi che la creazione della tabella fittizia sia registrata nei registri di archivio. A tale scopo, eseguire i seguenti comandi:

```
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. Per copiare l'ultimo dei registri di archivio, eseguire i seguenti comandi. Il database deve essere disponibile ma non aperto.

```
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. Per copiare i log di archivio, eseguire i seguenti comandi:

```
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. Infine, riprodurre i log di archivio rimanenti sul nuovo server.

```
[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 questa fase, replicare tutti i dati. Il database è pronto per essere convertito da un database di standby a un database operativo attivo e quindi aperto.

```
SQL> alter database activate standby database;
Database altered.
SQL> alter database open;
Database altered.
```

8. Verificare la presenza della tabella fittizia e poi rilasciarla.

```
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.
```

Migrazione dei log di ripristino senza interruzioni

A volte, un database è organizzato correttamente in generale, ad eccezione dei registri di ripristino. Questo può accadere per molte ragioni, la più comune delle quali è correlata agli snapshot. Prodotti come SnapManager per Oracle, SnapCenter e il framework di gestione dello storage NetApp Snap Creator consentono il ripristino quasi istantaneo di un database, ma solo se vengono ripristinati i volumi dei file di dati. Se i log di redo condividono lo spazio con i file di dati, non è possibile eseguire la reversione in modo sicuro, poiché causerebbe la distruzione dei log di redo, probabilmente la perdita di dati. Pertanto, i log di ripristino devono essere spostati.

Questa procedura è semplice e può essere eseguita senza interruzioni.

Configurazione corrente del log di ripristino

1. Identificare il numero di gruppi di log di ripristino e i rispettivi numeri di gruppo.

```
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. Immettere le dimensioni dei registri di ripristino.

```
SQL> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
--------------------------------------------------------------------------------
--------
1 524288000
2 524288000
3 524288000
```

Creare nuovi registri

1. Per ogni log di ripristino, creare un nuovo gruppo con dimensioni e numero di membri corrispondenti.

```
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. Verificare la nuova configurazione.

```
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.
```

Rilasciare i vecchi registri

1. Rilasciare i vecchi registri (gruppi 1, 2 e 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. Se si verifica un errore che impedisce di rilasciare un registro attivo, forzare un passaggio al registro successivo per rilasciare il blocco e forzare un checkpoint globale. Fare riferimento al seguente esempio di questo processo. Il tentativo di rilasciare il gruppo di file di registro 2, che si trovava nella vecchia posizione, è stato negato perché in questo file di registro erano ancora presenti dati attivi.

```
SQL> alter database drop logfile group 2;
alter database drop logfile group 2
*
ERROR at line 1:
ORA-01623: log 2 is current log for instance NTAP (thread 1) - cannot
drop
ORA-00312: online log 2 thread 1: '/redo0/NTAP/redo02a.log'
ORA-00312: online log 2 thread 1: '/redo1/NTAP/redo02b.log'
```

3. Un’archiviazione dei log seguita da un punto di verifica consente di rilasciare il file di log.

```
SQL> alter system archive log current;
System altered.
SQL> alter system checkpoint;
System altered.
SQL> alter database drop logfile group 2;
Database altered.
```

4. Quindi, eliminare i log dal file system. Questo processo deve essere eseguito con estrema attenzione.

Copia dei dati dell’host

Come per la migrazione a livello di database, la migrazione nel layer host fornisce un approccio indipendente dal vendor di soluzioni di storage.

In altre parole, talvolta "basta copiare i file" è l’opzione migliore.

Sebbene questo approccio a bassa tecnologia possa sembrare troppo semplice, offre comunque vantaggi significativi in quanto non è richiesto alcun software speciale e i dati originali rimangono intatti in tutta sicurezza durante il processo. Il limite principale è rappresentato dal fatto che la migrazione dei dati di una copia file causa interruzioni, poiché il database deve essere arrestato prima dell’inizio dell’operazione di copia. Non esiste un buon modo per sincronizzare le modifiche all’interno di un file, quindi i file devono essere completamente disattivati prima che la copia abbia inizio.

Se l’arresto richiesto da un’operazione di copia non è desiderabile, l’opzione successiva migliore basata su host è sfruttare un Logical Volume Manager (LVM). Esistono molte opzioni LVM, tra cui Oracle ASM, tutte con funzionalità simili, ma anche con alcune limitazioni che è necessario tenere in considerazione. Nella maggior parte dei casi, la migrazione può essere eseguita senza downtime e interruzioni.

Copia da filesystem a filesystem

L’utilità di una semplice operazione di copia non deve essere sottovalutata. Si tratta di un processo altamente affidabile che non richiede particolari competenze su sistemi operativi, database o sistemi storage. Inoltre, è molto sicuro perché non influisce sui dati originali. In genere, un amministratore di sistema modifica i file system di origine in modo che vengano montati in sola lettura e quindi riavvia un server per garantire che nessun elemento possa danneggiare i dati correnti. Il processo di copia può essere eseguito tramite script per garantire che venga eseguito il più rapidamente possibile senza il rischio di errori dell’utente. Poiché il tipo di i/o è un semplice trasferimento sequenziale dei dati, risulta estremamente efficiente in termini di larghezza di banda.

Nell’esempio seguente viene illustrata un’opzione per una migrazione sicura e rapida.

Ambiente

L’ambiente da migrare è il seguente:

- File system attuali

```
ontap-nfs1:/host1_oradata       52428800  16196928  36231872  31%
/oradata
ontap-nfs1:/host1_logs          49807360    548032  49259328   2% /logs
```

- Nuovi file system

```
ontap-nfs1:/host1_logs_new      49807360       128  49807232   1%
/new/logs
ontap-nfs1:/host1_oradata_new   49807360       128  49807232   1%
/new/oradata
```

Panoramica

Un DBA può migrare il database chiudendo semplicemente il database e copiando i file. Tuttavia, se occorre migrare molti database o ridurre al minimo il downtime, il processo può essere facilmente gestito tramite script. L'utilizzo di script riduce inoltre la possibilità di errori da parte dell'utente.

Gli script di esempio illustrati automatizzano le seguenti operazioni:

- Chiusura del database in corso
- Conversione dei file system esistenti in uno stato di sola lettura
- Copiare tutti i dati dai file system di origine a quelli di destinazione, mantenendo tutte le autorizzazioni dei file
- Smontaggio dei file system vecchi e nuovi
- Rimontaggio dei nuovi file system negli stessi percorsi dei file system precedenti

Procedura

1. Arrestare il database.

```
[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. Convertire i file system in sola lettura. Questa operazione può essere eseguita più rapidamente utilizzando uno script, come illustrato nella "Convertire il file system in sola lettura".

```
[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. Verificare che i file system siano ora di sola lettura.

```
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. Sincronizzare il contenuto del file system con `rsync` comando.

```
[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. Smontare i vecchi file system e riposizionare i dati copiati. Questa operazione può essere eseguita più rapidamente utilizzando uno script, come illustrato nella "Sostituire il file system".

```
[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. Verificare che i nuovi file system siano in posizione.

```
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. Avviare il database.

```
[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
```

Cutover completamente automatizzato

Questo script di esempio accetta argomenti del SID del database seguiti da coppie di file system delimitate in comune. Per l'esempio sopra illustrato, il comando viene inviato come segue:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
```

Quando viene eseguito, lo script di esempio tenta di eseguire la seguente sequenza. Termina se incontra un errore in qualsiasi fase:

1. Arrestare il database.
2. Convertire i file system correnti in stato di sola lettura.
3. Utilizzare ciascuna coppia di argomenti del file system delimitati da virgole e sincronizzare il primo file system con il secondo.
4. Smontare i file system precedenti.
5. Aggiornare `/etc/fstab` archiviare come segue:
   a. Creare un backup in `/etc/fstab.bak`.
   b. Annotare le voci precedenti per i file system precedenti e nuovi.
   c. Creare una nuova voce per il nuovo file system che utilizza il vecchio punto di montaggio.
6. Montare i file system.
7. Avviare il database.

Il testo seguente fornisce un esempio di esecuzione per questo script:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:05:50 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
sending incremental file list
```

```
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%  185.40MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   81.34MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   70.42MB/s    0:00:00 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   47.08MB/s    0:00:01 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
sent 527098156 bytes  received 278 bytes  150599552.57 bytes/sec
total size is 527032832  speedup is 1.00
Succesfully replicated filesystem /logs to /new/logs
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
 10737426432 100%  176.55MB/s    0:00:58 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%    9.48MB/s    0:00:02 (xfer#2, to-check=9/13)
... NTAP/undotbs01.dbf
   309338112 100%   70.76MB/s    0:00:04 (xfer#9, to-check=2/13)
NTAP/undotbs02.dbf
  1073750016 100%  187.65MB/s    0:00:05 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    5.09MB/s    0:00:00 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
```

```
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  177725933.55 bytes/sec
total size is 18570092218  speedup is 1.00
Succesfully replicated filesystem /oradata to /new/oradata
swap 0 /logs /new/logs
/new/logs unmounted
/logs unmounted
Mounted updated /logs
Swapped filesystem /logs for /new/logs
swap 1 /oradata /new/oradata
/new/oradata unmounted
/oradata unmounted
Mounted updated /oradata
Swapped filesystem /oradata for /new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:08:59 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
[root@host1 current]#
```

Migrazione Oracle ASM spfile e passwd

Una difficoltà nel completare la migrazione che coinvolge ASM è rappresentata dallo spfile specifico per ASM e dal file delle password. Per impostazione predefinita, questi file di metadati critici vengono creati nel primo gruppo di dischi ASM definito. Se un particolare gruppo di dischi ASM deve essere evacuato e rimosso, il file spfile e la password che governano l'istanza ASM deve essere riposizionato.

Un altro caso d'utilizzo in cui potrebbe essere necessario trasferire questi file è durante una distribuzione di software di gestione del database, come SnapManager per Oracle o il plug-in SnapCenter Oracle. Una delle funzionalità di questi prodotti è il ripristino rapido di un database ripristinando lo stato dei LUN ASM che ospitano i file di dati. Per eseguire questa operazione, è necessario portare il gruppo di dischi ASM offline prima di eseguire un ripristino. Questo non è un problema, purché i file di dati di un determinato database siano isolati in un gruppo di dischi ASM dedicato.

Quando il gruppo di dischi contiene anche il file ASM spfile/passwd, l'unico modo per mettere il gruppo di dischi in modalità non in linea è arrestare l'intera istanza ASM. Si tratta di un processo di interruzione, il che significa che il file spfile/passwd dovrebbe essere riposizionato.

Ambiente

1. SID database = TOAST
2. File di dati correnti su `+DATA`
3. File di log e file di controllo correnti attivati `+LOGS`
4. Nuovi gruppi di dischi ASM stabiliti come `+NEWDATA` e. `+NEWLOGS`

Posizioni dei file spfile/passwd ASM

Il trasferimento di questi file può essere eseguito senza interruzione delle attività. Tuttavia, per motivi di sicurezza, NetApp consiglia di arrestare l'ambiente del database in modo da poter essere certi che i file siano stati spostati e che la configurazione sia stata aggiornata correttamente. Questa procedura deve essere ripetuta se su un server sono presenti più istanze ASM.

Identificare le istanze ASM

Identificare le istanze ASM in base ai dati registrati in `oratab` file. Le istanze di ASM sono indicate dal simbolo +.

```
-bash-4.1$ cat /etc/oratab | grep '^+'
+ASM:/orabin/grid:N             # line added by Agent
```

Su questo server è presente un'istanza ASM denominata +ASM.

Assicurarsi che tutti i database siano chiusi

L'unico processo di smon visibile dovrebbe essere quello per l'istanza ASM in uso. La presenza di un altro processo di smon indica che un database è ancora in esecuzione.

```
-bash-4.1$ ps -ef | grep smon
oracle     857     1  0 18:26 ?        00:00:00 asm_smon_+ASM
```

L'unico processo di smon è l'istanza ASM stessa. Ciò significa che nessun altro database è in esecuzione ed è sicuro procedere senza il rischio di interrompere le operazioni del database.

Individuare i file

Identificare la posizione corrente del file spfile e della password di ASM utilizzando `spget` e. `pwget` comandi.

```
bash-4.1$ asmcmd
ASMCMD> spget
+DATA/spfile.ora
```

```
ASMCMD> pwget --asm
+DATA/orapwasm
```

I file si trovano entrambi alla base di `+DATA` gruppo di dischi.

Copiare i file

Copiare i file nel nuovo gruppo di dischi ASM con `spcopy` e. `pwcopy` comandi. Se il nuovo gruppo di dischi è stato creato di recente ed è attualmente vuoto, potrebbe essere necessario montarlo per primo.

```
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
```

I file sono stati copiati da `+DATA` a. `+NEWDATA`.

Aggiornare l'istanza ASM

L'istanza ASM deve ora essere aggiornata per riflettere la modifica della posizione. Il `spset` e. `pwset` I comandi aggiornano i metadati ASM richiesti per l'avvio del gruppo di dischi ASM.

```
ASMCMD> spset +NEWDATA/spfile.ora
ASMCMD> pwset --asm +NEWDATA/orapwasm
```

Attivare ASM utilizzando i file aggiornati

A questo punto, l'istanza ASM utilizza ancora le posizioni precedenti di questi file. L'istanza deve essere riavviata per forzare una rilettura dei file dalle nuove posizioni e per rilasciare i blocchi sui file precedenti.

```
-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
```

Rimuovere i vecchi file spfile e password

Se la procedura è stata eseguita correttamente, i file precedenti non sono più bloccati e possono essere rimossi.

```
-bash-4.1$ asmcmd
ASMCMD> rm +DATA/spfile.ora
ASMCMD> rm +DATA/orapwasm
```

Copia da Oracle ASM a ASM

Oracle ASM è essenzialmente un volume manager e un file system combinati e leggeri. Poiché il file system non è facilmente visibile, è necessario utilizzare RMAN per eseguire operazioni di copia. Sebbene il processo di migrazione basato sulle copie sia sicuro e semplice, si traduce in un'interruzione. È possibile ridurre al minimo le interruzioni, ma non eliminarle completamente.

Se si desidera eseguire la migrazione senza interruzioni di un database basato su ASM, l'opzione migliore è sfruttare la capacità di ASM di riequilibrare le estensioni ASM nei nuovi LUN, eliminando al contempo i vecchi LUN. In genere, questo tipo di operazioni è sicuro e senza interruzioni, ma non offre alcun percorso di back-out. Se si riscontrano problemi di funzionamento o di prestazioni, l'unica opzione è quella di trasferire nuovamente i dati all'origine.

Questo rischio può essere evitato copiando il database nella nuova posizione piuttosto che spostare i dati, in modo che i dati originali non vengano toccati. Il database può essere completamente testato nella sua nuova posizione prima di entrare in funzione e il database originale è disponibile come opzione di fallback se vengono rilevati problemi.

Questa procedura è una delle numerose opzioni che interessano RMAN. È progettato per consentire un processo in due fasi in cui viene creato il backup iniziale e quindi sincronizzato successivamente tramite la riproduzione del registro. Questo processo è auspicabile per ridurre al minimo i tempi di inattività, in quanto consente al database di rimanere operativo e di distribuire i dati durante la copia di base iniziale.

Copia database

Oracle RMAN crea una copia di livello 0 (completa) del database di origine attualmente presente nel gruppo di dischi ASM `+DATA` alla nuova posizione su `+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
```

Forzare l'interruttore del registro di archiviazione

È necessario forzare un'opzione del log di archivio per assicurarsi che i log di archivio contengano tutti i dati necessari per rendere la copia completamente coerente. Senza questo comando, i dati chiave potrebbero essere ancora presenti nei log di ripristino.

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Arrestare il database di origine

L'interruzione inizia in questa fase perché il database viene arrestato e inserito in una modalità di sola lettura ad accesso limitato. Per arrestare il database di origine, eseguire i seguenti comandi:

```
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 ControlFile

È necessario eseguire il backup di controlfile nel caso in cui sia necessario interrompere la migrazione e ripristinare la posizione di archiviazione originale. Una copia del controlfile di backup non è richiesta al 100%, ma rende più semplice il processo di ripristino delle posizioni dei file di database nella posizione originale.

```
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
```

Aggiornamenti dei parametri

Il file spfile corrente contiene riferimenti ai file di controllo nelle posizioni correnti all'interno del vecchio gruppo di dischi ASM. Deve essere modificato, il che è fatto facilmente modificando una versione pfile intermedia.

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Aggiornare pfile

Aggiornare tutti i parametri che fanno riferimento ai vecchi gruppi di dischi ASM per riflettere i nuovi nomi dei gruppi di dischi ASM. Quindi salvare il file pfile aggiornato. Assicurarsi che il `db_create` parametri presenti.

Nell'esempio seguente, i riferimenti a. `+DATA` che sono stati modificati in `+NEWDATA` sono evidenziati in giallo. Due parametri chiave sono `db_create` parametri che creano nuovi file nella posizione corretta.

```
*.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'
```

Aggiorna il file init.ora

La maggior parte dei database basati su ASM utilizza un `init.ora` file che si trova in `$ORACLE_HOME/dbs` Directory, che è un punto di spfile sul gruppo di dischi ASM. Questo file deve essere reindirizzato a una posizione sul nuovo gruppo di dischi ASM.

```
-bash-4.1$ cd $ORACLE_HOME/dbs
-bash-4.1$ cat initTOAST.ora
SPFILE='+DATA/TOAST/spfileTOAST.ora'
```

Modificare questo file come segue:

```
SPFILE=+NEWLOGS/TOAST/spfileTOAST.ora
```

Ricreazione del file dei parametri

spfile è ora pronto per essere popolato dai dati nel pfile modificato.

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Avviare il database per iniziare a utilizzare il nuovo spfile

Avviare il database per assicurarsi che utilizzi ora il nuovo spfile creato e che eventuali ulteriori modifiche ai parametri di sistema siano registrate correttamente.

```
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
```

Ripristina controlfile

Il controlfile di backup creato da RMAN può anche essere ripristinato da RMAN direttamente nella posizione specificata nel nuovo spfile.

```
RMAN> restore controlfile from
'+DATA/TOAST/CONTROLFILE/current.258.897683139';
Starting restore at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=417 device type=DISK
channel ORA_DISK_1: copied control file copy
output file name=+NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
Finished restore at 06-DEC-15
```

Montare il database e verificare l'uso del nuovo controlfile.

```
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
```

Riproduzione del registro

Il database utilizza attualmente i file di dati nella vecchia posizione. Prima di poter utilizzare la copia, è necessario sincronizzarla. È trascorso del tempo durante il processo di copia iniziale e le modifiche sono state registrate principalmente nei registri di archivio. Queste modifiche vengono replicate come segue:

1. Eseguire un backup incrementale RMAN, che contiene i registri di archivio.

```
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. Riprodurre nuovamente il registro.

```
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
```

Attivazione

Il controlfile ripristinato fa ancora riferimento ai file di dati nella posizione originale e contiene anche le informazioni di percorso per i file di dati copiati.

1. Per modificare i file di dati attivi, eseguire `switch database to copy` comando.

```
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"
```

   I file di dati attivi sono ora i file di dati copiati, ma potrebbero comunque essere presenti modifiche nei log di ripristino finali.

2. Per riprodurre tutti i registri rimanenti, eseguire il `recover database` comando. Se il messaggio `media recovery complete` il processo è stato eseguito correttamente.

```
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
```

Questo processo ha modificato solo la posizione dei file di dati normali. I file di dati temporanei devono essere rinominati, ma non devono essere copiati perché sono solo temporanei. Il database è attualmente inattivo, pertanto non sono presenti dati attivi nei file di dati temporanei.

3. Per spostare i file di dati temporanei, identificarne prima la posizione.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
--------------------------------------------------------------------------------
--------
1 +DATA/TOAST/TEMPFILE/temp.263.897683145
```

4. Spostare i file di dati temporanei utilizzando un comando RMAN che imposta il nuovo nome per ciascun file di dati. Con Oracle Managed Files (OMF), il nome completo non è necessario; il gruppo di dischi ASM è sufficiente. Quando il database viene aperto, OMF si collega alla posizione appropriata nel gruppo di dischi ASM. Per spostare i file, eseguire i seguenti comandi:

```
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
```

Migrazione dei log di ripristino

Il processo di migrazione è quasi completo, ma i log di ripristino si trovano ancora nel gruppo di dischi ASM originale. I log di ripristino non possono essere spostati direttamente. Viene invece creata una nuova serie di log di ripristino che viene aggiunta alla configurazione, seguita da una rimozione dei log precedenti.

1. Identificare il numero di gruppi di log di ripristino e i rispettivi numeri di gruppo.

```
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. Immettere le dimensioni dei registri di ripristino.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
--------------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Per ogni log di ripristino, creare un nuovo gruppo con una configurazione corrispondente. Se non si utilizza OMF, è necessario specificare il percorso completo. Questo è anche un esempio che utilizza `db_create_online_log` parametri. Come mostrato in precedenza, questo parametro era impostato su +NEWLOGS. Questa configurazione consente di utilizzare i seguenti comandi per creare nuovi registri online senza dover specificare un percorso di file o un gruppo di dischi ASM specifico.

```
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. Aprire il database.

```
SQL> alter database open;
Database altered.
```

5. Rilasciare i vecchi registri.

```
RMAN> alter database drop logfile group 1;
Statement processed
```

6. Se si verifica un errore che impedisce di rilasciare un registro attivo, forzare un passaggio al registro

successivo per rilasciare il blocco e forzare un checkpoint globale. Di seguito è riportato un esempio. Il tentativo di rilasciare il gruppo di file di registro 3, che si trovava nella vecchia posizione, è stato negato perché in questo file di registro erano ancora presenti dati attivi. L'archiviazione di un registro dopo un punto di verifica consente di eliminare il file di registro.

```
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. Esaminare l'ambiente per assicurarsi che tutti i parametri basati sulla posizione siano aggiornati.

```
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. Nello script seguente viene illustrato come semplificare questo processo:

```
[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. Se i gruppi di dischi ASM sono stati completamente evacuati, è possibile smontarli con `asmcmd`. Tuttavia, in molti casi i file appartenenti ad altri database o al file ASM spfile/passwd potrebbero essere ancora presenti.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Copia da Oracle ASM al file system

La procedura di copia da Oracle ASM a file system è molto simile alla procedura di copia da ASM a ASM, con vantaggi e restrizioni simili. La differenza principale è la sintassi dei vari comandi e parametri di configurazione quando si utilizza un file system visibile anziché un gruppo di dischi ASM.

Copia database

Oracle RMAN viene utilizzato per creare una copia di livello 0 (completa) del database di origine attualmente presente nel gruppo di dischi ASM `+DATA` alla nuova posizione su `/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
```

Forzare l’interruttore del registro di archiviazione

È necessario forzare lo switch del log di archivio per assicurarsi che i log di archivio contengano tutti i dati necessari per rendere la copia completamente coerente. Senza questo comando, i dati chiave potrebbero essere ancora presenti nei log di ripristino. Per forzare un’opzione del log di archivio, eseguire il comando seguente:

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Arrestare il database di origine

L'interruzione inizia in questa fase perché il database viene arrestato e inserito in una modalità di sola lettura ad accesso limitato. Per arrestare il database di origine, eseguire i seguenti comandi:

```
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 ControlFile

Eseguire il backup dei file di controllo nel caso in cui sia necessario interrompere la migrazione e ripristinare la posizione di archiviazione originale. Una copia del controlfile di backup non è richiesta al 100%, ma rende più semplice il processo di ripristino delle posizioni dei file di database nella posizione originale.

```
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
```

Aggiornamenti dei parametri

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Aggiornare pfile

Tutti i parametri che fanno riferimento ai vecchi gruppi di dischi ASM devono essere aggiornati e, in alcuni casi, eliminati quando non sono più rilevanti. Aggiornarli per riflettere i nuovi percorsi del file system e salvare il file pfile aggiornato. Assicurarsi che sia elencato il percorso di destinazione completo. Per aggiornare questi parametri, eseguire i seguenti comandi:

```
*.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'
```

Disattivare il file init.ora originale

Questo file si trova in `$ORACLE_HOME/dbs` Ed è in genere in un pfile che funge da puntatore a spfile sul gruppo di dischi ASM. Per assicurarsi che spfile originale non sia più utilizzato, rinominarlo. Non eliminarlo, tuttavia, perché questo file è necessario se la migrazione deve essere interrotta.

```
[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]$
```

Ricreazione del file dei parametri

Questa è la fase finale del trasferimento di spfile. Il file spfile originale non viene più utilizzato e il database viene avviato (ma non montato) utilizzando il file intermedio. Il contenuto di questo file può essere scritto nella nuova posizione spfile come segue:

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Avviare il database per iniziare a utilizzare il nuovo spfile

È necessario avviare il database per rilasciare i blocchi sul file intermedio e avviare il database utilizzando solo il nuovo file spfile. L'avvio del database dimostra inoltre che la nuova posizione di spfile è corretta e che i suoi dati sono validi.

```
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
```

Ripristina controlfile

È stato creato un controlfile di backup nel percorso `/tmp/TOAST.ctrl` nelle fasi precedenti della procedura. Il nuovo spfile definisce le posizioni controlfile come /`logfs/TOAST/ctrl/ctrlfile1.ctrl` e. `/logfs/TOAST/redo/ctrlfile2.ctrl`. Tuttavia, tali file non esistono ancora.

1. Questo comando ripristina i dati controlfile nei percorsi definiti in spfile.

   ```
   RMAN> restore controlfile from '/tmp/TOAST.ctrl';
   Starting restore at 13-MAY-16
   using channel ORA_DISK_1
   channel ORA_DISK_1: copied control file copy
   output file name=/logs/TOAST/arch/control01.ctl
   output file name=/logs/TOAST/redo/control02.ctl
   Finished restore at 13-MAY-16
   ```

2. Eseguire il comando mount in modo che i file di controllo vengano rilevati correttamente e contengano dati validi.

   ```
   RMAN> alter database mount;
   Statement processed
   released channel: ORA_DISK_1
   ```

   Per convalidare `control_files` eseguire il seguente comando:

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
/logs/TOAST/arch/control01.ctl
                                                 ,
/logs/TOAST/redo/control02.c
                                                 tl
```

Riproduzione del registro

Il database sta attualmente utilizzando i file di dati nella vecchia posizione. Prima di poter utilizzare la copia, è necessario sincronizzare i file di dati. È trascorso del tempo durante il processo di copia iniziale e le modifiche sono state registrate principalmente nei registri di archivio. Queste modifiche vengono replicate nei due passaggi seguenti.

1. Eseguire un backup incrementale RMAN, che contiene i registri di archivio.

```
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. Riprodurre i registri.

```
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
```

Attivazione

Il controlfile ripristinato fa ancora riferimento ai file di dati nella posizione originale e contiene anche le informazioni di percorso per i file di dati copiati.

1. Per modificare i file di dati attivi, eseguire `switch database to copy` comando:

```
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. Sebbene i file di dati debbano essere completamente coerenti, è necessario eseguire un passaggio finale per riprodurre le modifiche rimanenti registrate nei registri di ripristino online. Utilizzare `recover database` comando per riprodurre queste modifiche e rendere la copia identica al 100% all'originale. Tuttavia, la copia non è ancora aperta.

```
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
```

Spostare i file di dati temporanei

1. Identificare la posizione dei file di dati temporanei ancora in uso sul gruppo di dischi originale.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
-------------------------------------------------------------------------
--------
1 +ASM0/TOAST/temp01.dbf
```

2. Per spostare i file di dati, eseguire i seguenti comandi. Se ci sono molti tempfile, utilizzare un editor di testo per creare il comando RMAN e quindi tagliarlo e incollarlo.

```
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
```

Migrazione dei log di ripristino

Il processo di migrazione è quasi completo, ma i log di ripristino si trovano ancora nel gruppo di dischi ASM originale. I log di ripristino non possono essere spostati direttamente. Al contrario, viene creata e aggiunta alla configurazione una nuova serie di log di ripristino, in seguito a una perdita dei vecchi log.

1. Identificare il numero di gruppi di log di ripristino e i rispettivi numeri di gruppo.

```
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. Immettere le dimensioni dei registri di ripristino.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Per ogni log di ripristino, creare un nuovo gruppo utilizzando le stesse dimensioni del gruppo di log di ripristino corrente utilizzando la nuova posizione del file system.

```
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. Rimuovere i vecchi gruppi di file di registro che si trovano ancora nell'archivio precedente.

```
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. Se si verifica un errore che blocca l'eliminazione di un registro attivo, forzare un passaggio al registro successivo per rilasciare il blocco e forzare un punto di verifica globale. Di seguito è riportato un esempio. Il tentativo di rilasciare il gruppo di file di registro 3, che si trovava nella vecchia posizione, è stato negato

perché in questo file di registro erano ancora presenti dati attivi. L'archiviazione dei log seguita da un punto di verifica consente l'eliminazione dei file di log.

```
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. Esaminare l'ambiente per assicurarsi che tutti i parametri basati sulla posizione siano aggiornati.

```
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. Nel seguente script viene illustrato come semplificare questo processo.

```
[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. Se i gruppi di dischi ASM sono stati completamente evacuati, è possibile smontarli con `asmcmd`. In molti casi, i file appartenenti ad altri database o al file ASM spfile/passwd possono essere ancora presenti.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Procedura di pulizia del file di dati

Il processo di migrazione potrebbe generare file di dati con sintassi lunga o criptica, a seconda del modo in cui è stato utilizzato Oracle RMAN. Nell'esempio illustrato, il backup è stato eseguito con il formato file di `/oradata/TOAST/%U`. `%U` Indica che RMAN deve creare un nome univoco predefinito per ciascun file di dati. Il risultato è simile a quanto illustrato nel testo seguente. I nomi tradizionali dei file di dati sono incorporati nei nomi. Questo può essere ripulito utilizzando l'approccio basato su script illustrato nella "Pulitura della migrazione ASM".

```
[root@jfsc1 current]# ./fixuniquenames.pl TOAST
#sqlplus Commands
shutdown immediate;
startup mount;
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/system.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/sysaux.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt /oradata/TOAST/undotbs1.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
/oradata/TOAST/users.dbf
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSTEM_FNO-1_01r5fhjg' to '/oradata/TOAST/system.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSAUX_FNO-2_02r5fhjo' to '/oradata/TOAST/sysaux.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
UNDOTBS1_FNO-3_03r5fhjt' to '/oradata/TOAST/undotbs1.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
USERS_FNO-4_05r5fhk6' to '/oradata/TOAST/users.dbf';
alter database open;
```

Ribilanciamento di Oracle ASM

Come indicato in precedenza, è possibile eseguire la migrazione trasparente di un gruppo di dischi Oracle ASM in un nuovo sistema di storage utilizzando il processo di ribilanciamento. Riassumendo, il processo di ribilanciamento richiede l'aggiunta di LUN di dimensioni uguali al gruppo esistente di LUN, seguita da un'operazione di disgregazione del LUN precedente. Oracle ASM riposiziona automaticamente i dati sottostanti nel nuovo storage in un layout ottimale e, al termine, rilascia i vecchi LUN.

Il processo di migrazione utilizza un i/o sequenziale efficiente e non causa generalmente un'interruzione delle performance, ma la velocità di migrazione può essere rallentata quando necessario.

Identificazione dei dati da migrare

```
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
```

Creazione di nuovi LUN

Creare nuovi LUN delle stesse dimensioni e impostare l'appartenenza a utenti e gruppi come richiesto. I LUN devono essere visualizzati come `CANDIDATE` dischi.

```
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
```

Aggiungere nuovi LUN

Anche se è possibile eseguire tutte le operazioni di aggiunta e rilascio, in genere è più semplice aggiungere nuovi LUN in due passaggi. Innanzitutto, aggiungere i nuovi LUN al gruppo di dischi. Questo passaggio comporta la migrazione di metà delle estensioni dai LUN ASM correnti ai nuovi LUN.

La potenza di riequilibrio indica la velocità di trasferimento dei dati. Più alto è il numero, più alto è il parallelismo del trasferimento dei dati. La migrazione viene eseguita con efficienti operazioni di i/o sequenziali che hanno scarse probabilità di causare problemi di performance. Tuttavia, se lo si desidera, il potere di riequilibrio di una migrazione in corso può essere regolato con `alter diskgroup [name] rebalance power [level]` comando. Le migrazioni tipiche utilizzano un valore di 5.

```
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586b' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315869' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315858' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586a' rebalance power 5;
Diskgroup altered.
```

Funzionamento del monitor

È possibile monitorare e gestire un'operazione di ribilanciamento in più modi. Per questo esempio è stato utilizzato il comando seguente.

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Una volta completata la migrazione, non vengono segnalate operazioni di ribilanciamento.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

LUN meno recenti

La migrazione è ormai a metà strada. Potrebbe essere opportuno eseguire alcuni test delle prestazioni di base per assicurarsi che l'ambiente sia sano. Dopo la conferma, è possibile spostare i dati rimanenti eliminando i vecchi LUN. Tenere presente che ciò non determina il rilascio immediato dei LUN. L'operazione di rilascio indica ad Oracle ASM di riposizionare prima le estensioni e quindi rilasciare il LUN.

```
sqlplus / as sysasm
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0000 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0001 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0002 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0003 rebalance power 5;
Diskgroup altered.
```

Funzionamento del monitor

L'operazione di ribilanciamento può essere monitorata e gestita in più modi. Per questo esempio è stato utilizzato il seguente comando:

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Una volta completata la migrazione, non vengono segnalate operazioni di ribilanciamento.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Rimuovere i vecchi LUN

Prima di rimuovere i vecchi LUN dal gruppo di dischi, è necessario eseguire un controllo finale dello stato dell'intestazione. Dopo il rilascio di un LUN da ASM, non viene più elencato un nome e lo stato dell'intestazione viene elencato come `FORMER`. Questo indica che questi LUN possono essere rimossi in modo sicuro dal sistema.

```
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.
```

Migrazione LVM

La procedura qui presentata mostra i principi di una migrazione basata su LVM di un gruppo di volumi chiamato `datavg`. Gli esempi sono tratti da Linux LVM, ma i principi si applicano ugualmente a AIX, HP-UX e VxVM. I comandi precisi possono variare.

1. Identificare i LUN attualmente presenti in `datavg` gruppo di volumi.

```
[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. Creazione di nuovi LUN di dimensioni fisiche identiche o leggermente superiori e definizione di volumi fisici.

```
[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. Aggiungere i nuovi volumi al gruppo di volumi.

```
[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. Eseguire il `pvmove` Comando per spostare le estensioni di ogni LUN corrente nel nuovo LUN. Il `-i [seconds]` l'argomento controlla l'avanzamento dell'operazione.

```
[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. Una volta completato questo processo, rimuovere i LUN precedenti dal gruppo di volumi utilizzando `vgreduce` comando. Se l'operazione ha esito positivo, è ora possibile rimuovere il LUN dal sistema in modo sicuro.

```
[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"
```

Importazione LUN esterne

Pianificazione

Le procedure per migrare LE risorse SAN utilizzando FLI sono documentate in NetApp "Documentazione per l'importazione dei LUN esteri di ONTAP" .

Dal punto di vista del database e dell'host, non sono necessarie operazioni speciali. Dopo l'aggiornamento delle zone FC e la disponibilità dei LUN su ONTAP, LVM dovrebbe essere in grado di leggere i metadati LVM dai LUN. Inoltre, i gruppi di volumi sono pronti per l'uso senza ulteriori passaggi di configurazione. Rari casi, gli ambienti potrebbero includere file di configurazione con hard-code e riferimenti allo storage array precedente. Ad esempio, un sistema Linux che includeva `/etc/multipath.conf` Le regole che fanno riferimento a un WWN di un dato dispositivo devono essere aggiornate per riflettere le modifiche introdotte da FLI.

> Fare riferimento alla matrice di compatibilità NetApp per informazioni sulle configurazioni supportate. Se il proprio ambiente non è incluso, contattare il rappresentante NetApp per assistenza.

Questo esempio mostra la migrazione di LUN ASM e LVM ospitati su un server Linux. FLI è supportato su altri sistemi operativi e, sebbene i comandi sul lato host possano differire, i principi sono gli stessi e le procedure ONTAP sono identiche.

Identificare i LUN LVM

La prima fase della preparazione consiste nell'identificare i LUN da migrare. Nell'esempio mostrato qui, due file system basati su SAN sono montati su `/orabin` e. `/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
```

Il nome del gruppo di volumi può essere estratto dal nome del dispositivo, che utilizza il formato (nome del gruppo di volumi)-(nome del volume logico). In questo caso, viene chiamato il gruppo di volumi `sanvg`.

Il `pvdisplay` Il comando può essere utilizzato come segue per identificare i LUN che supportano questo gruppo di volumi. In questo caso, sono presenti 10 LUN che compongono il `sanvg` gruppo di volumi.

```
[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
```

Identificare i LUN ASM

Anche i LUN ASM devono essere migrati. Per ottenere il numero di LUN e percorsi LUN da sqlplus come utente sysasm, eseguire il comando seguente:

```
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>
```

Modifiche alla rete FC

L'ambiente corrente contiene 20 LUN da migrare. Aggiornare la SAN corrente in modo che ONTAP possa accedere ai LUN correnti. I dati non sono ancora stati migrati, ma ONTAP deve leggere le informazioni di configurazione dalle LUN correnti per creare la nuova home page per quei dati.

Almeno una porta HBA sul sistema AFF/FAS deve essere configurata come porta Initiator. Inoltre, le zone FC devono essere aggiornate in modo che ONTAP possa accedere alle LUN sullo storage array esterno. Alcuni storage array hanno configurato il masking dei LUN, che limita i WWN che possono accedere a una determinata LUN. In tal caso, è necessario aggiornare anche il masking dei LUN per garantire l'accesso ai WWN di ONTAP.

Al termine di questa operazione, ONTAP dovrebbe essere in grado di visualizzare l'array di archiviazione esterno con `storage array show` comando. Il campo chiave restituito è il prefisso utilizzato per identificare il LUN esterno sul sistema. Nell'esempio seguente, i LUN dell'array esterno `FOREIGN_1` Appare in ONTAP usando il prefisso di `FOR-1`.

Identificare un array esterno

```
Cluster01::> storage array show -fields name,prefix
name          prefix
------------- ------
FOREIGN_1     FOR-1
Cluster01::>
```

Identificare i LUN esterni

I LUN possono essere elencati passando l' `array-name` al `storage disk show` comando. I dati restituiti vengono referenziati più volte durante la procedura di migrazione.

```
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::>
```

Registrare LUN di array esterni come candidati di importazione

Le LUN esterne vengono inizialmente classificate come qualsiasi tipo di LUN specifico. Prima di poter importare i dati, i LUN devono essere contrassegnati come esterni e quindi come candidati al processo di importazione. Questo passaggio viene completato passando il numero di serie a. `storage disk modify`, come illustrato nell'esempio seguente. Si noti che questa procedura etichetta solo il LUN come estraneo all'interno di ONTAP. Nessun dato viene scritto nella LUN esterna stessa.

```
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::*>
```

Creazione di volumi per l'hosting di LUN migrati

Per ospitare le LUN migrate è necessario un volume. La configurazione esatta dei volumi dipende dal piano generale per sfruttare le funzionalità di ONTAP. In questo esempio, i LUN ASM vengono posizionati in un volume e i LUN LVM in un secondo volume. In questo modo, puoi gestire le LUN come gruppi indipendenti per scopi come il tiering, la creazione di snapshot o l'impostazione di controlli della qualità del servizio.

Impostare `snapshot-policy `to `none`. Il processo di migrazione può comportare un notevole ricambio dei dati. Pertanto, potrebbe verificarsi un notevole aumento del consumo di spazio se le istantanee vengono create accidentalmente perché i dati indesiderati vengono acquisiti nelle istantanee.

```
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::>
```

Creare LUN ONTAP

Una volta creati i volumi, è necessario creare i nuovi LUN. In genere, la creazione di un LUN richiede all'utente di specificare tali informazioni come la dimensione LUN, ma in questo caso l'argomento del disco esterno viene passato al comando. Di conseguenza, ONTAP replica i dati di configurazione LUN correnti dal numero di serie specificato. Utilizza inoltre la geometria del LUN e i dati della tabella delle partizioni per regolare l'allineamento delle LUN e stabilire prestazioni ottimali.

In questo passaggio, i numeri di serie devono essere referenziati rispetto all'array esterno per assicurarsi che il LUN esterno corretto corrisponda al nuovo LUN corretto.

```
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)
```

Creare relazioni di importazione

I LUN sono stati creati ma non sono configurati come destinazione di replica. Prima di eseguire questo passaggio, i LUN devono essere messi offline. Questo passaggio aggiuntivo è progettato per proteggere i dati dagli errori dell'utente. Se ONTAP consentisse di eseguire una migrazione su un LUN online, rischierebbe di provocare la sovrascrittura dei dati attivi con un errore tipografico. Questa fase aggiuntiva, che obbliga l'utente a portare un LUN offline, consente di verificare se viene utilizzato il LUN di destinazione corretto come destinazione della migrazione.

```
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
```

Una volta che i LUN sono offline, è possibile stabilire la relazione di importazione passando il numero di serie del LUN esterno a. `lun import create` comando.

```
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::*>
```

Una volta stabilite tutte le relazioni di importazione, è possibile riportare online i LUN.

```
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::*>
```

Crea gruppo iniziatore

Un gruppo iniziatore (igroup) fa parte dell'architettura di mascheramento LUN di ONTAP. Un LUN appena creato non è accessibile a meno che non venga concesso per la prima volta l'accesso a un host. A tale scopo, creare un igroup in cui siano elencati i nomi WWN FC o iSCSI Initiator a cui è necessario concedere l'accesso. Al momento della scrittura del report, FLI era supportato solo per LUN FC. Tuttavia, la conversione in post-migrazione iSCSI è un'attività semplice, come illustrato nella "Conversione protocollo".

In questo esempio, viene creato un igroup che contiene due WWN corrispondenti alle due porte disponibili sull'HBA dell'host.

```
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
```

Mappare nuovi LUN all'host

Dopo la creazione di igroup, i LUN vengono quindi mappati all'igroup definito. Questi LUN sono disponibili solo per i WWN inclusi in questo igroup. In questa fase del processo di migrazione, NetApp presume che l'host non sia stato sottoposto a zoning in ONTAP. Questo è importante perché se l'host è contemporaneamente collegato all'array esterno e al nuovo sistema ONTAP, vi è il rischio che su ogni array possano essere rilevati LUN con lo stesso numero di serie. Questa situazione potrebbe causare malfunzionamenti del multipath o danni ai dati.

```
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::*>
```

Cutover

Una parte delle interruzioni durante l'importazione di LUN esterne è inevitabile a causa della necessità di modificare la configurazione di rete FC. Tuttavia, l'interruzione non deve durare più a lungo del tempo necessario per riavviare l'ambiente di database e

aggiornare lo zoning FC per passare dalla connettività FC dell'host al ONTAP.

Questo processo può essere riassunto come segue:

1. Quiete di tutta l'attività LUN sui LUN esterni.
2. Reindirizzare le connessioni FC dell'host al nuovo sistema ONTAP.
3. Attivare il processo di importazione.
4. Rilevare nuovamente i LUN.
5. Riavviare il database.

Non è necessario attendere il completamento del processo di migrazione. Non appena inizia la migrazione di un determinato LUN, questo è disponibile su ONTAP e può fornire dati durante il processo di copia dei dati. Tutte le letture vengono passate alla LUN esterna e tutte le scritture vengono scritte in modo sincrono su entrambi gli array. L'operazione di copia è molto veloce e l'overhead del reindirizzamento del traffico FC è minimo, per cui qualsiasi impatto sulle performance deve essere transitorio e minimo. In caso di problemi, è possibile ritardare il riavvio dell'ambiente fino al completamento del processo di migrazione e all'eliminazione delle relazioni di importazione.

Chiudere il database

Il primo passo per chiudere l'ambiente in questo esempio è arrestare il database.

```
[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>
```

Chiudere i servizi di rete

Uno dei file system basati su SAN oggetto della migrazione include anche i servizi Oracle ASM. La disattivazione dei LUN sottostanti richiede lo smontaggio dei file system, il che a sua volta significa l'arresto di tutti i processi con file aperti su questo file system.

```
[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]$
```

Smontare i file system

Se tutti i processi vengono arrestati, l'operazione umount ha esito positivo. Se l'autorizzazione viene negata, è necessario che sul file system sia presente un processo con blocco. Il `fuser` command può aiutare a identificare questi processi.

```
[root@host1 ~]# umount /orabin
[root@host1 ~]# umount /backups
```

Disattivare i gruppi di volumi

Una volta smontati tutti i file system di un dato gruppo di volumi, è possibile disattivare il gruppo di volumi.

```
[root@host1 ~]# vgchange --activate n sanvg
  0 logical volume(s) in volume group "sanvg" now active
[root@host1 ~]#
```

Modifiche alla rete FC

È ora possibile aggiornare le zone FC per rimuovere tutti gli accessi dall'host all'array esterno e stabilire l'accesso a ONTAP.

Avviare il processo di importazione

Per avviare i processi di importazione LUN, eseguire `lun import start` comando.

```
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*>
```

Monitorare l'avanzamento dell'importazione

L'operazione di importazione può essere monitorata con `lun import show` comando. Come illustrato di seguito, è in corso l'importazione di tutte le LUN da 20 GB, il che significa che i dati sono ora accessibili tramite ONTAP, anche se l'operazione di copia dei dati continua a proseguire.

```
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.
```

Se è necessario un processo offline, ritardare il riscoperta o il riavvio dei servizi finché il comando non `lun import show` indica che la migrazione è stata completata correttamente. È quindi possibile completare il processo di migrazione come descritto in "Importazione di LUN esterne - completamento".

Se hai bisogno di una migrazione online, procedi con il rilevamento dei LUN nella nuova sede e attiva i servizi.

Eseguire la scansione delle modifiche al dispositivo SCSI

Nella maggior parte dei casi, l'opzione più semplice per ritrovare nuove LUN è riavviare l'host. In questo modo, si rimuovono automaticamente i vecchi dispositivi obsoleti, si rilevano correttamente tutti i nuovi LUN e si creano dispositivi associati come i dispositivi multipathing. L'esempio qui mostra una procedura completamente online a scopo dimostrativo.

Attenzione: Prima di riavviare un host, assicurarsi che tutte le voci in `/etc/fstab` Il riferimento alle risorse SAN migrate verrà commentato. Se questa operazione non viene eseguita e si verificano problemi con l'accesso LUN, il sistema operativo potrebbe non avviarsi. Questa situazione non danneggia i dati. Tuttavia, può essere molto scomodo avviare in modalità rescue o in una modalità simile e correggere `/etc/fstab` In modo che il sistema operativo possa essere avviato per consentire la risoluzione dei problemi.

I LUN della versione di Linux utilizzata in questo esempio possono essere rianalizzati con `rescan-scsi-bus.sh` comando. Se il comando viene eseguito correttamente, nell'output viene visualizzato ogni percorso LUN. L'output può essere difficile da interpretare, ma, se la configurazione di zoning e igroup era corretta, molti LUN dovrebbero apparire che includono un `NETAPP` stringa fornitore.

```
[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.
```

Verificare la presenza di dispositivi multipercorso

Il processo di rilevamento LUN attiva anche la ricreazione dei dispositivi multipath, ma è noto che il driver multipathing Linux presenta problemi occasionali. L'output di `multipath -ll` dovrebbe essere controllato per verificare che l'output sia come previsto. Per esempio, l'uscita seguente mostra dispositivi multipercorso associati a A. `NETAPP` stringa fornitore. Ciascun dispositivo dispone di quattro percorsi, di cui due con priorità 50 e due con priorità 10. Anche se l'output esatto può variare con diverse versioni di Linux, questo risultato

sembra come previsto.

Fare riferimento alla documentazione delle utilità host per la versione di Linux utilizzata per verificare che `/etc/multipath.conf` le impostazioni sono corrette.

```
[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
```

Riattivare il gruppo di volumi LVM

Se i LUN LVM sono stati rilevati correttamente, l' `vgchange --activate y` il comando dovrebbe riuscire. Questo è un buon esempio del valore di un volume manager logico. Una modifica del WWN di una LUN o anche di un numero di serie non è importante perché i metadati del gruppo di volumi vengono scritti sul LUN stesso.

Il sistema operativo ha eseguito la scansione dei LUN e ha rilevato una piccola quantità di dati scritti sul LUN che lo identifica come volume fisico appartenente a. `sanvg volumegroup`. Successivamente, ha costruito tutti i dispositivi necessari. È sufficiente riattivare il gruppo di volumi.

```
[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
```

Rimontare i file system

Dopo la riattivazione del gruppo di volumi, i file system possono essere montati con tutti i dati originali intatti. Come indicato in precedenza, i file system sono completamente operativi anche se la replica dei dati è ancora attiva nel gruppo back.

```
[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
```

Ripetere la scansione per i dispositivi ASM

I dispositivi ASMlib dovrebbero essere stati riselezionati al momento della nuova scansione dei dispositivi SCSI. La riscoperta può essere verificata online riavviando ASMlib e quindi eseguendo la scansione dei dischi.

Questa fase è pertinente solo alle configurazioni ASM in cui viene utilizzato ASMlib.

Attenzione: Se non viene utilizzato ASMlib, il `/dev/mapper` i dispositivi dovrebbero essere stati ricreati automaticamente. Tuttavia, le autorizzazioni potrebbero non essere corrette. È necessario impostare autorizzazioni speciali sui dispositivi sottostanti per ASM in assenza di ASMlib. Questa operazione viene solitamente eseguita tramite voci speciali in entrambi `/etc/multipath.conf` oppure `udev` o eventualmente in entrambi i set di regole. È possibile che questi file debbano essere aggiornati per riflettere le modifiche apportate all'ambiente in termini di numeri WWN o di serie per assicurarsi che i dispositivi ASM dispongano ancora delle autorizzazioni corrette.

In questo esempio, il riavvio di ASMlib e la scansione dei dischi mostrano gli stessi 10 LUN ASM dell'ambiente originale.

```
[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"
```

Riavviare i servizi di rete

Ora che i dispositivi LVM e ASM sono online e disponibili, è possibile riavviare i servizi grid.

```
[root@host1 /]# cd /orabin/product/12.1.0/grid/bin
[root@host1 bin]# ./crsctl start has
```

Riavviare il database

Dopo aver riavviato i servizi di griglia, è possibile avviare il database. Potrebbe essere necessario attendere alcuni minuti affinché i servizi ASM diventino completamente disponibili prima di provare ad avviare il database.

```
[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>
```

Completamento

Dal punto di vista dell'host, la migrazione è completa, ma l'i/o viene ancora servito dall'array esterno fino a quando le relazioni di importazione non vengono eliminate.

Prima di eliminare le relazioni, è necessario confermare che il processo di migrazione è completo per tutte le LUN.

```
Cluster01::*> lun import show -vserver vserver1 -fields foreign-
disk,path,operational-state
vserver   foreign-disk path              operational-state
--------- ------------ ----------------- -----------------
vserver1 800DT$HuVWB/ /vol/new_asm/LUN4 completed
vserver1 800DT$HuVWBW /vol/new_asm/LUN0 completed
vserver1 800DT$HuVWBX /vol/new_asm/LUN1 completed
vserver1 800DT$HuVWBY /vol/new_asm/LUN2 completed
vserver1 800DT$HuVWBZ /vol/new_asm/LUN3 completed
vserver1 800DT$HuVWBa /vol/new_asm/LUN5 completed
vserver1 800DT$HuVWBb /vol/new_asm/LUN6 completed
vserver1 800DT$HuVWBc /vol/new_asm/LUN7 completed
vserver1 800DT$HuVWBd /vol/new_asm/LUN8 completed
vserver1 800DT$HuVWBe /vol/new_asm/LUN9 completed
vserver1 800DT$HuVWBf /vol/new_lvm/LUN0 completed
vserver1 800DT$HuVWBg /vol/new_lvm/LUN1 completed
vserver1 800DT$HuVWBh /vol/new_lvm/LUN2 completed
vserver1 800DT$HuVWBi /vol/new_lvm/LUN3 completed
vserver1 800DT$HuVWBj /vol/new_lvm/LUN4 completed
vserver1 800DT$HuVWBk /vol/new_lvm/LUN5 completed
vserver1 800DT$HuVWBl /vol/new_lvm/LUN6 completed
vserver1 800DT$HuVWBm /vol/new_lvm/LUN7 completed
vserver1 800DT$HuVWBn /vol/new_lvm/LUN8 completed
vserver1 800DT$HuVWBo /vol/new_lvm/LUN9 completed
20 entries were displayed.
```

Elimina relazioni di importazione

Al termine del processo di migrazione, eliminare la relazione di migrazione. Dopo aver fatto ciò, l'i/o viene servito esclusivamente dalle unità su ONTAP.

```
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN9
```

Annullare la registrazione di LUN esterne

Infine, modificare il disco per rimuovere `is-foreign` designazione.

```
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::*>
```

Conversione del protocollo

La modifica del protocollo utilizzato per accedere a un LUN è un requisito comune.

In alcuni casi, fa parte di una strategia globale di migrazione dei dati nel cloud. TCP/IP è il protocollo del cloud e il passaggio da FC a iSCSI facilita la migrazione in vari ambienti cloud. In altri casi, iSCSI potrebbe essere desiderabile per sfruttare i costi ridotti di un IP SAN. A volte, una migrazione potrebbe utilizzare un protocollo diverso come misura temporanea. Ad esempio, se un array esterno e LUN basati su ONTAP non possono coesistere sugli stessi HBA, è possibile utilizzare LUN iSCSI abbastanza a lungo da copiare i dati dal vecchio array. Dopo la rimozione dei vecchi LUN dal sistema, è possibile riconvertirli in FC.

La seguente procedura illustra la conversione da FC a iSCSI, ma i principi generali si applicano a una conversione da iSCSI a FC inversa.

Installare iSCSI Initiator

La maggior parte dei sistemi operativi include un iniziatore iSCSI software per impostazione predefinita, ma se non è incluso, può essere facilmente installato.

```
[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 /]#
```

Identificare il nome dell'iniziatore iSCSI

Durante il processo di installazione viene generato un nome iSCSI initiator univoco. Su Linux, si trova in `/etc/iscsi/initiatorname.iscsi` file. Questo nome viene utilizzato per identificare l'host sulla SAN IP.

```
[root@host1 /]# cat /etc/iscsi/initiatorname.iscsi
InitiatorName=iqn.1992-05.com.redhat:497bd66ca0
```

Creare un nuovo gruppo iniziatore

Un gruppo iniziatore (igroup) fa parte dell'architettura di mascheramento LUN di ONTAP. Un LUN appena creato non è accessibile a meno che non venga concesso per la prima volta l'accesso a un host. Questa operazione viene eseguita creando un igroup che elenca i nomi WWN FC o iniziatori iSCSI che richiedono l'accesso.

In questo esempio, viene creato un igroup che contiene l'iniziatore iSCSI dell'host Linux.

```
Cluster01::*> igroup create -igroup linuxiscsi -protocol iscsi -ostype
linux -initiator iqn.1994-05.com.redhat:497bd66ca0
```

Chiudere l'ambiente

Prima di modificare il protocollo LUN, è necessario disattivare completamente i LUN. Tutti i database di uno dei LUN da convertire devono essere chiusi, i file system devono essere dismontati e i gruppi di volumi devono essere disattivati. Se si utilizza ASM, assicurarsi che il gruppo di dischi ASM sia smontato e chiudere tutti i servizi della griglia.

Rimuovere la mappatura dei LUN dalla rete FC

Una volta terminate completamente le LUN, rimuovere le mappature dall'igroup FC originale.

```
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
```

Eseguire nuovamente il mapping dei LUN alla rete IP

Concedere l'accesso a ogni LUN al nuovo gruppo di iniziatori basati su iSCSI.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxiscsi
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxiscsi
Cluster01::*>
```

Rilevamento delle destinazioni iSCSI

Il rilevamento iSCSI richiede due fasi. La prima è scoprire le destinazioni, che non è la stessa cosa per scoprire un LUN. Il `iscsiadm` il comando mostrato di seguito verifica il gruppo di portali specificato da `-p argument` Memorizza un elenco di tutti gli indirizzi IP e le porte che offrono servizi iSCSI. In questo caso, vi sono quattro indirizzi IP con servizi iSCSI sulla porta predefinita 3260.

> Il completamento di questo comando può richiedere alcuni minuti se non è possibile raggiungere uno qualsiasi degli indirizzi IP di destinazione.

```
[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
```

Rilevamento delle LUN iSCSI

Dopo aver rilevato le destinazioni iSCSI, riavviare il servizio iSCSI per rilevare i LUN iSCSI disponibili e creare i dispositivi associati, ad esempio i dispositivi multipath o ASMlib.

```
[root@host1 ~]# service iscsi restart
Redirecting to /bin/systemctl restart  iscsi.service
```

Riavviare l'ambiente

Riavviare l'ambiente riattivando i gruppi di volumi, rimontando i file system, riavviando i servizi RAC e così via. Per precauzione, NetApp consiglia di riavviare il server al termine del processo di conversione, per assicurarsi che tutti i file di configurazione siano corretti e che tutti i dispositivi obsoleti vengano rimossi.

Attenzione: Prima di riavviare un host, assicurarsi che tutte le voci in `/etc/fstab` Il riferimento alle risorse SAN migrate verrà commentato. Se questa operazione non viene eseguita e si verificano problemi con l'accesso LUN, il risultato può essere un sistema operativo che non si avvia. Questo problema non danneggia i dati. Tuttavia, può essere molto scomodo avviare in modalità rescue o una modalità simile e corretta `/etc/fstab` In modo che il sistema operativo possa essere avviato per consentire l'avvio delle operazioni di risoluzione dei problemi.

Script di esempio

Gli script presentati sono forniti come esempi di come eseguire lo script di varie attività del sistema operativo e del database. Vengono forniti così come sono. Se è necessario supporto per una procedura particolare, contattare NetApp o un rivenditore NetApp.

Arresto del database

Lo script Perl seguente prende un singolo argomento del SID Oracle e chiude un database. Può essere eseguito come utente Oracle o come root.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
77 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF4
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
`;};
print @out;
if ("@out" =~ /ORACLE instance shut down/) {
print "$oraclesid shut down\n";
exit 0;}
elsif ("@out" =~ /Connected to an idle instance/) {
print "$oraclesid already shut down\n";
exit 0;}
else {
print "$oraclesid failed to shut down\n";
exit 1;}
```

Avvio del database

Lo script Perl seguente prende un singolo argomento del SID Oracle e chiude un database. Può essere eseguito come utente Oracle o come root.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
'
`;}
else {
@out=`. oraenv << EOF3
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
`;};
print @out;
if ("@out" =~ /Database opened/) {
print "$oraclesid started\n";
exit 0;}
elsif ("@out" =~ /cannot start already-running ORACLE/) {
print "$oraclesid already started\n";
exit 1;}
else {
78 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
print "$oraclesid failed to start\n";
exit 1;}
```

Convertire il file system in sola lettura

Lo script seguente prende un argomento del file system e tenta di smontarlo e rimontarlo in modalità di sola lettura. Questa operazione è utile durante i processi di migrazione in cui un file system deve essere mantenuto disponibile per replicare i dati e deve essere protetto contro danni accidentali.

```
#! /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;
```

Sostituire il file system

L'esempio di script riportato di seguito viene utilizzato per sostituire un file system con un altro. Poiché modifica il file ``/etc/fstab, deve essere eseguito come root. Accetta un singolo argomento delimitato da virgole per i file system vecchi e nuovi.

1. Per sostituire il file system, eseguire lo script seguente:

```
#! /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;
```

Come esempio di utilizzo di questo script, si supponga che i dati in `/oradata` viene migrato in `/neworadata` e. `/logs` viene migrato in `/newlogs`. Uno dei metodi più semplici per eseguire questa attività consiste nell'utilizzare una semplice operazione di copia dei file per riportare la nuova periferica al punto di montaggio originale.

2. Si supponga che i file system vecchi e nuovi siano presenti in `/etc/fstab` archiviare come segue:

```
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. Quando viene eseguito, questo script smonta il file system corrente e lo sostituisce con il nuovo:

```
[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. Lo script aggiorna anche `/etc/fstab` file di conseguenza. Nell'esempio illustrato, sono incluse le seguenti modifiche:

```
#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
```

Migrazione automatizzata del database

In questo esempio viene illustrato l'utilizzo di script di arresto, avvio e sostituzione del file system per automatizzare completamente la migrazione.

```
#! /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;
```

Visualizzare le posizioni dei file

Questo script raccoglie una serie di parametri critici del database e li stampa in un formato di facile lettura. Questo script può essere utile quando si esaminano i layout dei dati. Inoltre, lo script può essere modificato per altri usi.

```
#! /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";
```

Pulitura della migrazione ASM

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
```

```
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;}
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print @out;
print "shutdown immediate;\n";
print "startup mount;\n";
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "host mv $line $1$newname\n";}}
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "alter database rename file '$line' to
'$1$newname';\n";}}
print "alter database open;\n";
print "\n";
```

Conversione del nome da ASM a file system

```
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;
/
```

Riprodurre i log sul database

Questo script accetta un singolo argomento di un SID Oracle per un database in modalità mount e tenta di riprodurre tutti i log di archivio attualmente disponibili.

```
#! /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;
```

Riprodurre i registri sul database di standby

Questo script è identico allo script precedente, tranne che è progettato per un database di standby.

```
#! /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;
```

Note aggiuntive

Ottimizzazione delle prestazioni e benchmarking

Il test accurato delle performance dello storage del database è un argomento estremamente complicato. Richiede la comprensione dei seguenti problemi:

- IOPS e throughput
- La differenza tra le operazioni i/o in primo piano e in background
- L'effetto della latenza sul database
- Numerose impostazioni del sistema operativo e di rete che influiscono sulle performance dello storage

Inoltre, occorre prendere in considerazione attività che non riguardano i database di storage. Esiste un punto in cui l'ottimizzazione delle performance dello storage non produce vantaggi utili perché le performance dello storage non sono più un fattore limitante per le performance.

La maggior parte dei clienti che utilizzano database sceglie ora gli array all-flash, il che crea alcune

considerazioni aggiuntive. Ad esempio, prendi in considerazione il test delle performance su un sistema AFF A900 a due nodi:

- Con un rapporto di lettura/scrittura di 80/20:1, due nodi A900 possono fornire oltre 1M IOPS di database casuali prima che la latenza attraversi anche il contrassegno 150µs. Questo ben oltre le attuali richieste di performance della maggior parte dei database è difficile prevedere il miglioramento previsto. Lo storage verrebbe ampiamente cancellato come collo di bottiglia.
- La larghezza di banda della rete è una fonte sempre più comune di limitazioni delle prestazioni. Ad esempio, le soluzioni su disco a rotazione sono spesso dei colli di bottiglia per le performance dei database perché la latenza i/o è molto elevata. Quando un array all-flash rimuove le limitazioni di latenza, spesso la barriera passa alla rete. Si tratta di un aspetto particolarmente interessante nel caso di ambienti virtualizzati e sistemi blade in cui è difficile visualizzare la vera connettività di rete. Ciò può complicare il test delle performance se il sistema di storage stesso non può essere pienamente utilizzato a causa di limitazioni della larghezza di banda.
- Generalmente, il confronto delle performance di un array all-flash con un array contenente dischi rotanti non è possibile a causa dell'aumento drastico della latenza degli array all-flash. I risultati dei test in genere non sono significativi.
- Il confronto delle performance di picco degli IOPS con un array all-flash spesso non è un test utile, in quanto i database non sono limitati dall'i/o dello storage Ad esempio, si supponga che un array sia in grado di sostenere 500K IOPS casuali, mentre un altro possa sostenere 300K KB. La differenza è irrilevante nel mondo reale se un database impiega il 99% del suo tempo per l'elaborazione della CPU. I carichi di lavoro non utilizzano mai le funzionalità complete dello storage array. Al contrario, le funzionalità degli IOPS di picco potrebbero essere critiche in una piattaforma di consolidamento in cui si prevede che lo storage array venga caricato alle proprie funzionalità di picco.
- In qualsiasi test dello storage, si tiene sempre in considerazione sia la latenza che gli IOPS. Molti storage array sul mercato dichiarano livelli estremi di IOPS, ma la latenza rende quegli IOPS inutili a tali livelli. La destinazione tipica degli array all-flash è il contrassegno 1ms. Un approccio migliore al test non consiste nel misurare gli IOPS massimi possibili, ma nel determinare quanti IOPS può supportare uno storage array prima che la latenza media sia superiore a 1ms ms.

Oracle Automatic workload Repository e benchmarking

Il gold standard per i confronti delle performance Oracle è un report Oracle Automatic workload Repository (AWR).

Esistono diversi tipi di rapporti AWR. Da un punto di vista dello storage, un report generato dall'esecuzione di `awrrpt.sql` È il comando più completo e utile, in quanto è destinato a una specifica istanza del database e include alcuni istogrammi dettagliati che suddividono gli eventi i/o dello storage in base alla latenza.

Il confronto fra due array delle performance implica l'esecuzione idealmente dello stesso carico di lavoro su ciascun array e la produzione di un report AWR che punta esattamente al carico di lavoro. Nel caso di un carico di lavoro con esecuzione molto lunga, è possibile utilizzare un singolo rapporto AWR con un tempo trascorso che comprende il tempo di inizio e di fine, ma è preferibile suddividere i dati AWR come rapporti multipli. Ad esempio, se un processo batch è stato eseguito dalla mezzanotte alle 6, creare una serie di rapporti AWR di un'ora dalle 1:1 alle 2:00 e così via.

In altri casi, è necessario ottimizzare una query molto breve. L'opzione migliore è un report AWR basato su uno snapshot AWR creato all'inizio della query e un secondo snapshot AWR creato al termine della query. Il server di database dovrebbe essere altrimenti silenzioso per ridurre al minimo l'attività in background che potrebbe oscurare l'attività della query in analisi.

Laddove i report AWR non sono disponibili, i report statspack Oracle sono una buona alternativa. Contengono la maggior parte delle stesse statistiche i/o di un rapporto AWR.

Oracle AWR e risoluzione dei problemi

Un report AWR è anche lo strumento più importante per analizzare un problema di prestazioni.

Come per il benchmarking, il troubleshooting delle performance richiede la misurazione precisa di un determinato carico di lavoro. Quando possibile, fornisci dati AWR quando segnali un problema di performance al centro di supporto NetApp o quando lavori con un account team NetApp o partner in merito a una nuova soluzione.

Quando si forniscono i dati AWR, considerare i seguenti requisiti:

- Eseguire `awrrpt.sql` per generare il report. L'output può essere di testo o HTML.
- Se si utilizzano Oracle Real Application Clusters (RAC), generare report AWR per ciascuna istanza del cluster.
- Indicare l'ora specifica in cui si è verificato il problema. Il tempo massimo accettabile trascorso di un rapporto AWR è generalmente di un'ora. Se un problema persiste per più ore o richiede un'operazione multi-ora, ad esempio un processo batch, fornire più rapporti AWR di un'ora che coprono l'intero periodo da analizzare.
- Se possibile, regolare l'intervallo dell'istantanea AWR su 15 minuti. Questa impostazione consente di eseguire un'analisi più dettagliata. Ciò richiede anche ulteriori esecuzioni di `awrrpt.sql` per fornire un report per ogni intervallo di 15 minuti.
- Se il problema è una query in esecuzione molto breve, fornire un report AWR basato su uno snapshot AWR creato all'inizio dell'operazione e un secondo snapshot AWR creato al termine dell'operazione. Il server di database dovrebbe essere altrimenti silenzioso per ridurre al minimo l'attività in background che oscurrebbe l'attività dell'operazione in analisi.
- Se viene segnalato un problema di prestazioni in determinati momenti ma non in altri, fornire dati AWR aggiuntivi che dimostrino buone prestazioni per il confronto.

calibra_io

Il `calibrate_io` command non deve mai essere utilizzato per testare, confrontare o eseguire il benchmark dei sistemi storage. Come indicato nella documentazione di Oracle, questa procedura calibra le funzionalità i/o dello storage.

La calibrazione non è la stessa del benchmarking. Lo scopo di questo comando è di emettere i/o per aiutare a calibrare le operazioni di database e migliorarne l'efficienza ottimizzando il livello di i/o inviato all'host. Poiché il tipo di i/o eseguito da `calibrate_io` L'operazione non rappresenta l'i/o effettivo dell'utente del database, i risultati non sono prevedibili e spesso non sono nemmeno riproducibili.

SLOB2

SLOB2, il Silly Little Oracle Benchmark, è diventato lo strumento preferito per la valutazione delle prestazioni del database. È stato sviluppato da Kevin Closson ed è disponibile su "https://kevinclosson.net/slob/". Occorrono pochi minuti per installare e configurare, oltre a utilizzare un database Oracle effettivo per generare schemi di i/o su una tablespace definibile dall'utente. È una delle poche opzioni di test disponibili in grado di saturare un array all-flash con l'i/O. È utile anche per generare livelli molto inferiori di i/o per simulare carichi di lavoro di storage che sono IOPS bassi ma sensibili alla latenza.

Panca di rotazione

Swingbench può essere utile per testare le prestazioni del database, ma è estremamente difficile utilizzare Swingbench in un modo che mette a dura prova lo storage. NetApp non ha riscontrato test da Swingbench che hanno dato i/o sufficienti per essere un carico significativo su qualsiasi array AFF. In casi limitati, è possibile utilizzare Order Entry Test (OET) per valutare lo storage dal punto di vista della latenza. Ciò può essere utile in situazioni in cui un database ha una dipendenza di latenza nota per determinate query. Assicurarsi che l'host e la rete siano configurati correttamente per realizzare i potenziali di latenza di un array all-flash.

HammerDB

HammerDB è uno strumento di test del database che simula, tra gli altri, i benchmark TPC-C e TPC-H. La creazione di un set di dati di dimensioni sufficienti per eseguire correttamente un test può richiedere molto tempo, ma può rivelarsi uno strumento efficace per valutare le prestazioni delle applicazioni OLTP e di data warehouse.

Orion

Lo strumento Oracle Orion è stato comunemente utilizzato con Oracle 9, ma non è stato mantenuto per garantire la compatibilità con le modifiche in vari sistemi operativi host. Viene raramente utilizzato con Oracle 10 o Oracle 11 a causa di incompatibilità con il sistema operativo e la configurazione dello storage.

Oracle ha riscritto lo strumento e viene installato per impostazione predefinita con Oracle 12c. Sebbene questo prodotto sia stato migliorato e utilizzi molte delle stesse chiamate utilizzate da un database Oracle reale, non utilizza esattamente lo stesso percorso di codice o lo stesso comportamento i/o utilizzato da Oracle. Ad esempio, la maggior parte degli i/o Oracle viene eseguita in modo sincrono, il che significa che il database si arresta finché l'i/o non viene completato quando l'operazione i/o viene completata in primo piano. Il semplice flooding di un sistema storage con i/o casuali non rappresenta una riproduzione di i/o Oracle reali e non offre un metodo diretto per confrontare gli array di storage o misurare l'effetto delle modifiche alla configurazione.

Detto questo, ci sono alcuni casi d'utilizzo per Orion, come la misurazione generale delle massime prestazioni possibili di una particolare configurazione host-rete-storage, o per misurare lo stato di un sistema storage. Con un test accurato, è possibile ideare test Orion utilizzabili per confrontare gli storage array o valutare l'effetto di una modifica della configurazione, a condizione che i parametri includano la considerazione di IOPS, throughput e latenza e cercare di replicare fedelmente un carico di lavoro realistico.

NFSv3 serrature obsolete

Se un server di database Oracle si blocca, potrebbe essersi verificato un problema con blocchi NFS obsoleti al riavvio. Questo problema può essere evitato prestando particolare attenzione alla configurazione della risoluzione dei nomi sul server.

Questo problema si verifica perché la creazione di un blocco e la cancellazione di un blocco utilizzano due metodi di risoluzione dei nomi leggermente diversi. Sono coinvolti due processi: Network Lock Manager (NLM) e il client NFS. NLM utilizza `uname n` per determinare il nome host, mentre la `rpc.statd` usi di processo `gethostbyname()`. Questi nomi host devono corrispondere affinché il sistema operativo elimini correttamente i blocchi obsoleti. Ad esempio, l'host potrebbe cercare i blocchi di proprietà di `dbserver5`, ma i blocchi sono stati registrati dall'host come `dbserver5.mydomain.org`. Se `gethostbyname()` non restituisce lo stesso valore di `uname -a`, quindi il processo di rilascio del blocco non ha avuto esito positivo.

Il seguente script di esempio verifica se la risoluzione dei nomi è completamente coerente:

```
#! /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";
```

Se `gethostbyname` non corrisponde `uname`, è probabile che siano presenti blocchi obsoleti. Ad esempio, questo risultato rivela un potenziale problema:

```
uname -n yields: dbserver5
gethostbyname yields: dbserver5.mydomain.org
```

La soluzione viene generalmente trovata modificando l'ordine in cui gli host vengono visualizzati `/etc/hosts`. Ad esempio, si supponga che il file hosts includa questa voce:

```
10.156.110.201  dbserver5.mydomain.org dbserver5 loghost
```

Per risolvere il problema, modificare l'ordine di visualizzazione del nome di dominio completo e del nome host breve:

```
10.156.110.201  dbserver5 dbserver5.mydomain.org loghost
```

`gethostbyname()` ora restituisce il breve `dbserver5` nome host, che corrisponde all'output di `uname`. I blocchi vengono quindi cancellati automaticamente dopo un arresto anomalo del server.

Verifica dell'allineamento di WAFL

Il corretto allineamento dell'WAFL è fondamentale per garantire buone prestazioni. Sebbene ONTAP gestisca blocchi in 4KB unità, questo fatto non significa che ONTAP esegua tutte le operazioni in 4KB unità. Infatti, ONTAP supporta operazioni a blocchi di diverse dimensioni, ma la contabilità sottostante è gestita da WAFL in 4KB unità.

Il termine "allineamento" si riferisce al modo in cui l'i/o Oracle corrisponde a queste unità 4KB. Per ottenere prestazioni ottimali è necessario che un blocco Oracle 8KB risieda su due blocchi fisici da 4KB WAFL su un'unità. Se un blocco è sfalsato di 2KB, questo blocco risiede su metà di un blocco 4KB, un blocco 4KB completo separato e quindi sulla metà di un terzo blocco 4KB. Questa disposizione causa un peggioramento delle prestazioni.

L'allineamento non è un problema con i file system NAS. I file di dati Oracle sono allineati all'inizio del file in base alle dimensioni del blocco Oracle. Pertanto, le dimensioni dei blocchi di 8KB, 16KB e 32KB sono sempre allineate. Tutte le operazioni di blocco sono sfalsate dall'inizio del file in unità di 4 kilobyte.

I LUN, al contrario, contengono generalmente qualche tipo di intestazione del driver o metadati del file system all'inizio che creano un offset. L'allineamento è raramente un problema nei sistemi operativi moderni, perché

questi sistemi operativi sono progettati per unità fisiche che potrebbero utilizzare un settore 4KB nativo, che richiede anche l'allineamento dell'i/o ai confini del 4KB per ottenere prestazioni ottimali.

Ci sono, tuttavia, alcune eccezioni. È possibile che un database sia stato migrato da un sistema operativo meno recente non ottimizzato per i/o 4KB o che un errore utente durante la creazione della partizione abbia causato un offset che non è in unità di 4KB.

I seguenti esempi sono specifici per Linux, ma la procedura può essere adattata per qualsiasi sistema operativo.

Allineato

L'esempio seguente mostra un controllo dell'allineamento su un singolo LUN con una singola partizione.

Innanzitutto, creare la partizione che utilizza tutte le partizioni disponibili sul disco.

```
[root@host0 iscsi]# fdisk /dev/sdb
Device contains neither a valid DOS partition table, nor Sun, SGI or OSF
disklabel
Building a new DOS disklabel with disk identifier 0xb97f94c1.
Changes will remain in memory only, until you decide to write them.
After that, of course, the previous content won't be recoverable.
The device presents a logical sector size that is smaller than
the physical sector size. Aligning to a physical sector (or optimal
I/O) size boundary is recommended, or performance may be impacted.
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-10240, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-10240, default 10240):
Using default value 10240
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
[root@host0 iscsi]#
```

L'allineamento può essere controllato matematicamente con il seguente comando:

```
[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
```

L'output mostra che le unità sono 512 byte, e l'inizio della partizione è 32 unità. Si tratta di un totale di 32 x 512 = 16.834 byte, ovvero un multiplo intero di 4KB blocchi WAFL. Questa partizione è allineata correttamente.

Per verificare il corretto allineamento, attenersi alla seguente procedura:

1. Identificare l'UUID (Universal Unique Identifier) del LUN.

```
FAS8040SAP::> lun show -v /vol/jfs_luns/lun0
              Vserver Name: jfs
                  LUN UUID: ed95d953-1560-4f74-9006-85b352f58fcd
                    Mapped: mapped`                `
```

2. Immettere la shell del nodo sul controller ONTAP.

```
FAS8040SAP::> node run -node FAS8040SAP-02
Type 'exit' or 'Ctrl-D' to return to the CLI
FAS8040SAP-02> set advanced
set not found.  Type '?' for a list of commands
FAS8040SAP-02> priv set advanced
Warning: These advanced commands are potentially dangerous; use
         them only when directed to do so by NetApp
         personnel.
```

3. Avviare le raccolte statistiche sull'UUID di destinazione identificato nel primo passaggio.

```
FAS8040SAP-02*> stats start lun:ed95d953-1560-4f74-9006-85b352f58fcd
Stats identifier name is 'Ind0xffffff08b9536188'
FAS8040SAP-02*>
```

4. Eseguire alcuni i/O. È importante utilizzare `iflag` Argomento per assicurarsi che i/o sia sincrono e non bufferizzato.

Prestare molta attenzione con questo comando. Inversione del `if` e. `of` gli argomenti distruggono i dati.

```
[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. Arrestare le statistiche e visualizzare l'istogramma di allineamento. Tutti i i/o devono trovarsi in `.0` Bucket, che indica i/o allineato al limite di un blocco 4KB.

```
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%
```

Disallineato

L'esempio seguente mostra i/o disallineati:

1. Creare una partizione che non si allinea a un confine 4KB. Questo non è il comportamento predefinito sui sistemi operativi moderni.

```
[root@host0 iscsi]# fdisk -u /dev/sdb
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (32-20971519, default 32): 33
Last sector, +sectors or +size{K,M,G} (33-20971519, default 20971519):
Using default value 20971519
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

2. La partizione è stata creata con un offset a 33 settori anziché con il valore predefinito 32. Ripetere la procedura descritta in "Allineato". L'istogramma viene visualizzato come segue:

```
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%
```

Il disallineamento è chiaro. L'i/o rientra principalmente in* *`.1` benna, che corrisponde all'offset previsto. Quando la partizione è stata creata, è stata spostata di 512 byte più avanti nel dispositivo rispetto al valore predefinito ottimizzato, il che significa che l'istogramma è spostato di 512 byte.

Inoltre, il `read_partial_blocks` Le statistiche sono diverse da zero, il che significa che è stato eseguito l'i/o che non ha riempito l'intero blocco da 4KB KB.

Ripristina la logging

Le procedure qui spiegate sono applicabili ai file di dati. I log di ripristino e gli archivi di Oracle hanno modelli di i/o diversi. Ad esempio, il redo logging è una sovrascrittura circolare di un singolo file. Se si utilizza la dimensione predefinita del blocco da 512 byte, le statistiche di scrittura sono simili a queste:

```
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%
```

L'i/o viene distribuito in tutti i bucket di istogramma, ma non si tratta di un problema di prestazioni. Velocità di redo-logging estremamente elevate potrebbero, tuttavia, trarre vantaggio dall'utilizzo di dimensioni del blocco di 4KB KB. In questo caso, è consigliabile assicurarsi che i LUN di redo-logging siano allineati correttamente. Tuttavia, questo non è importante per le buone prestazioni come l'allineamento dei file dati.