Disaster recovery Oracle

Enterprise applications

NetApp
May 09, 2024

Sommario

Disaster recovery Oracle

Disaster recovery dei database Oracle con ONTAP

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.

SnapMirror Active Sync si basa su SnapMirror Synchronous. 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.

Sebbene la sincronizzazione attiva di SnapMirror e MetroCluster funzionino in modo molto diverso internamente, per un host il risultato è molto simile. La differenza principale è la granularità. Se hai solo bisogno di workload selezionati da replicare sincroni, l'opzione migliore è SnapMirror Active Sync. MetroCluster è l'opzione migliore per replicare interi ambienti o persino data center. Inoltre, la sincronizzazione attiva di SnapMirror è attualmente IMPOSTATA solo SU SAN, mentre MetroCluster è multiprotocollo, inclusi SAN, NFS e SMB.

MetroCluster

Architettura fisica di MetroCluster e database Oracle

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

[NOTA]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.

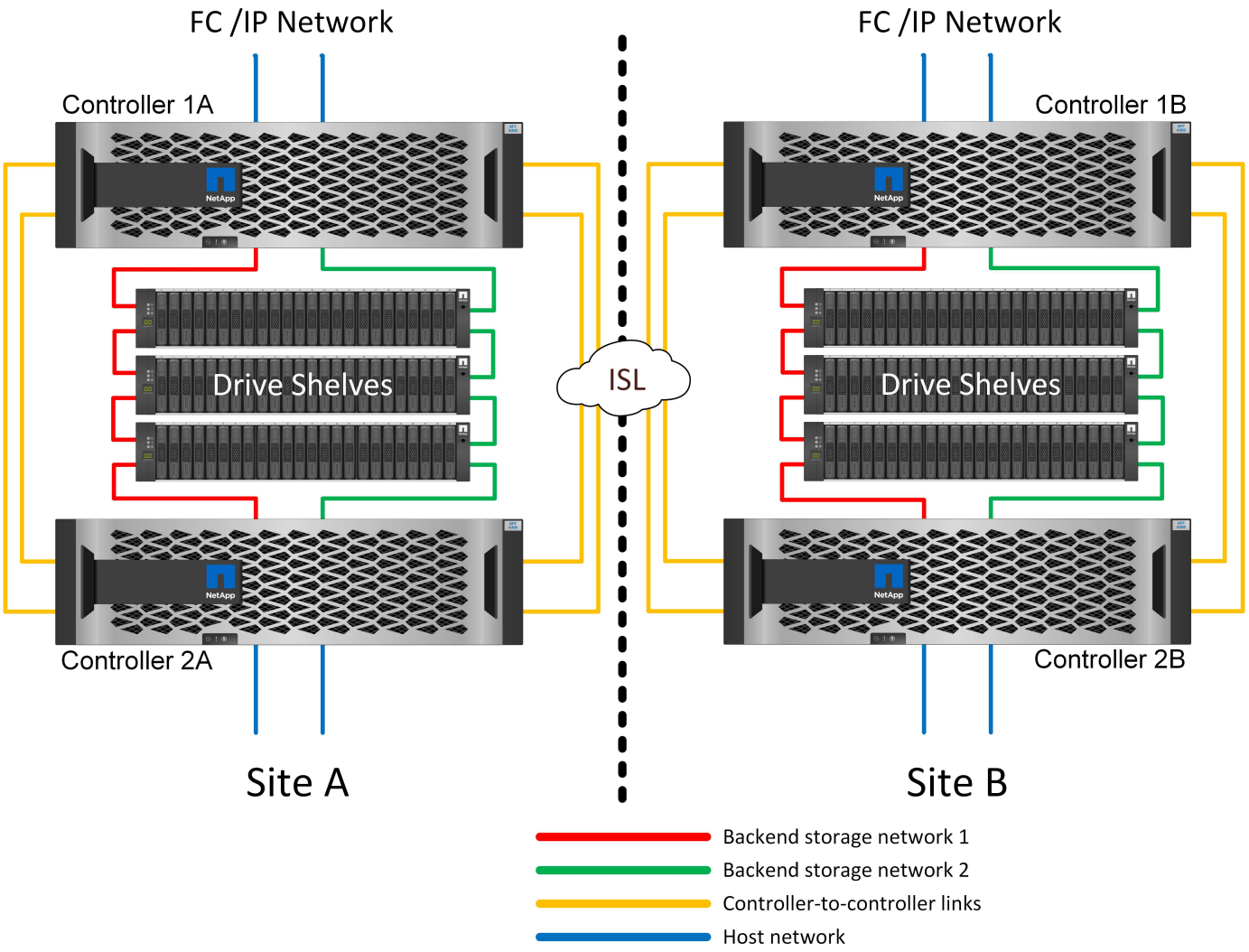
IP MetroCluster

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.

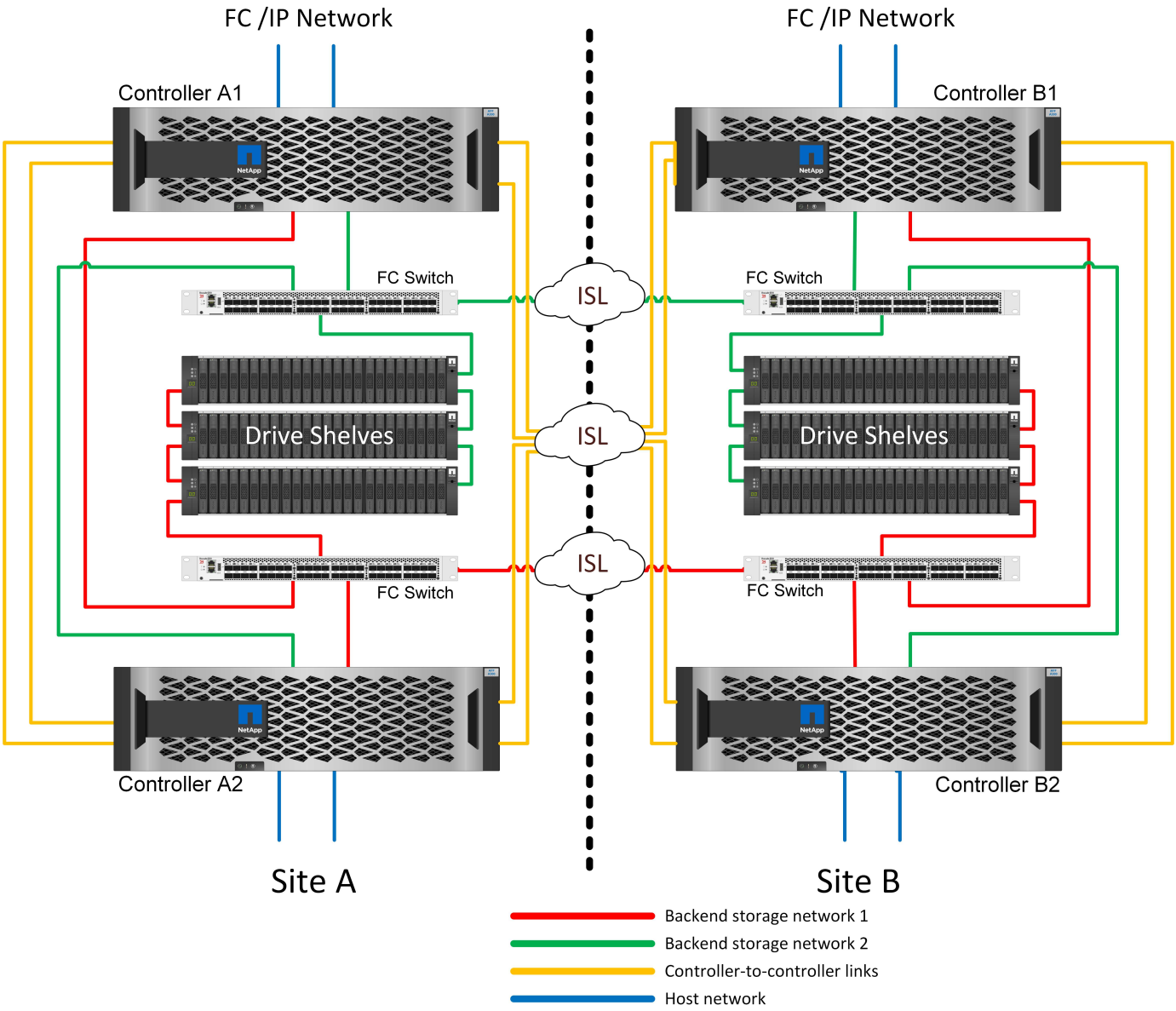
L'architettura è anche più semplice perché le uniche connessioni cross-site sono per i controller. Nei MetroClusters collegati a FC SAN, un controller scrive direttamente sulle unità del sito opposto e quindi richiede connessioni SAN, switch e bridge aggiuntivi. Al contrario, un controller in una configurazione 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.

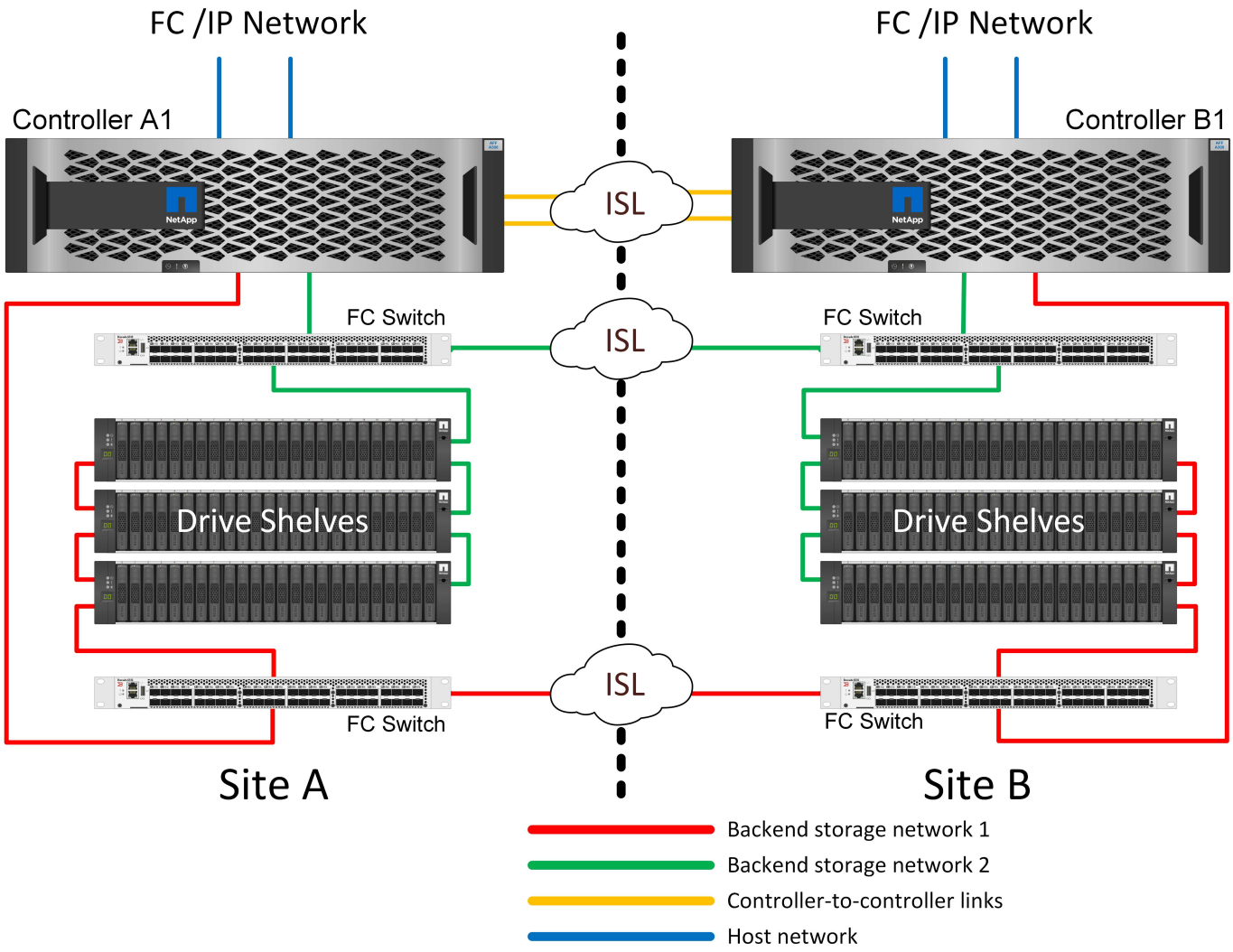


Alcune infrastrutture multisito non sono progettate per le operazioni Active-Active, ma vengono utilizzate maggiormente come sito primario e sito di disaster recovery. In questa situazione, è generalmente preferibile un'opzione ha-Pair MetroCluster per i seguenti motivi:

- 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 MetroCluster e database Oracle

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

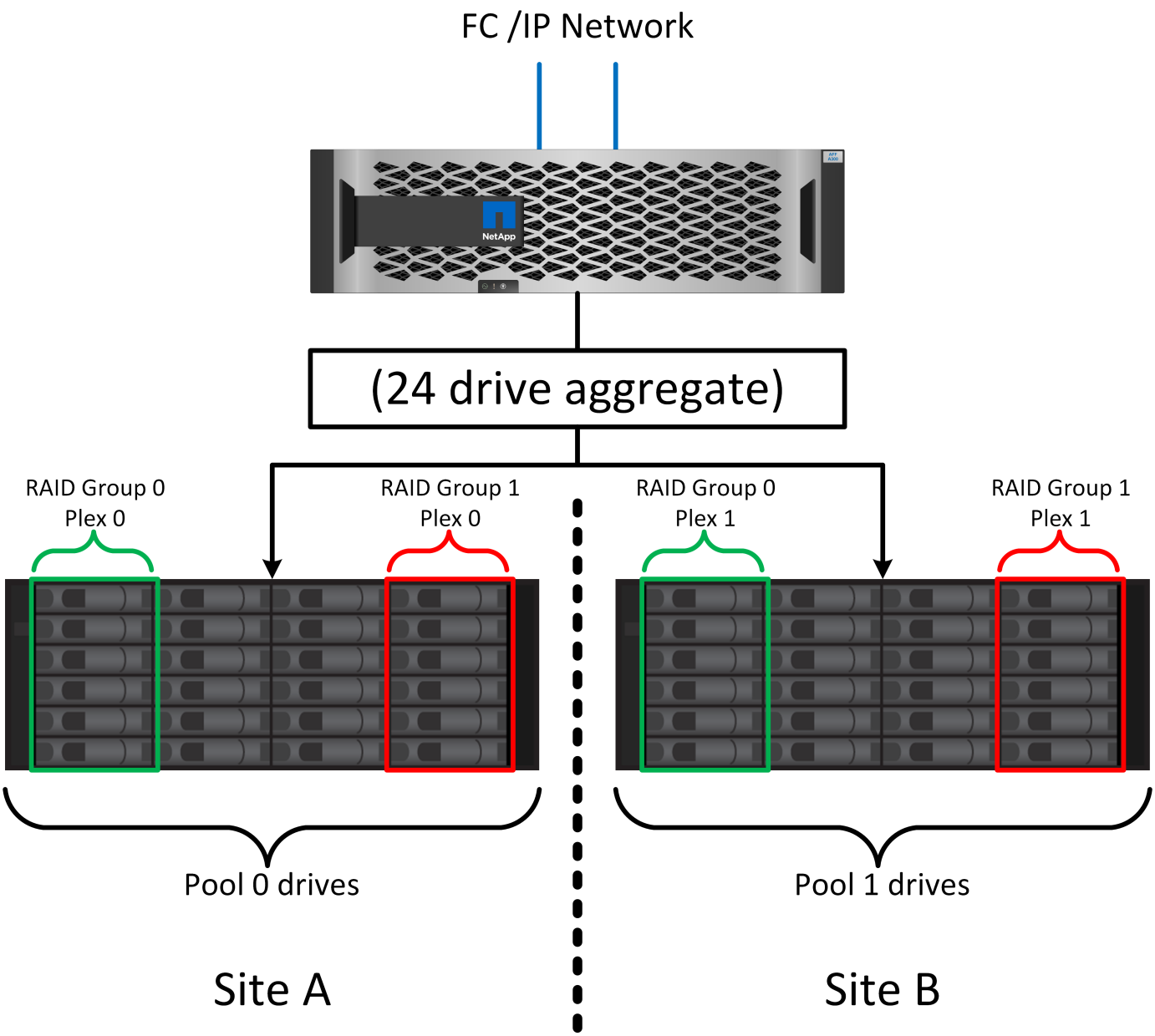
MetroCluster estende la protezione dei dati NVRAM nei seguenti modi:

- 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" È possibile eseguire il software 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 del rompighiaccio è disponibile sul "Sito di supporto NetApp", 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.

*Attenzione: *Il software MCTB non verifica che la NVRAM sia e/o i plessi siano 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

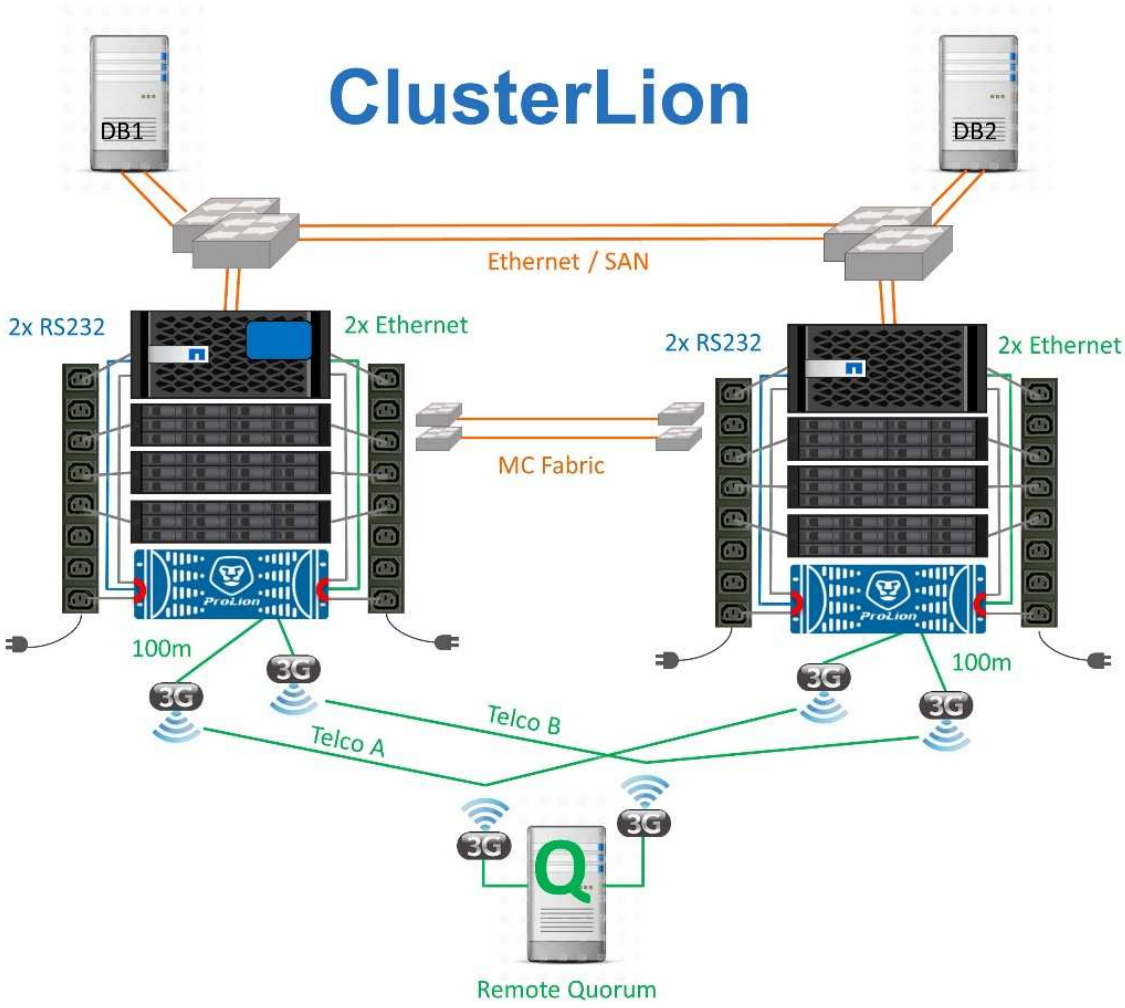
ONTAP Mediator viene utilizzato con MetroCluster IP e con alcune altre soluzioni ONTAP. Funziona come un servizio di tiebreaker tradizionale, proprio come il software MetroCluster Tiebreaker descritto in precedenza, ma include anche una funzione critica che consente di eseguire uno switchover automatizzato 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

Database Oracle con 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 timeout estesi per il mirroring del disco lato host. Per ulteriori informazioni e opzioni, consulta il tuo NetApp o il partner account team.

Failover del database Oracle con MetroCluster

```
 Metrocluster is an ONTAP feature that can protect your Oracle databases
with RPO=0 synchronous mirroring across sites, and it scales up to support
hundreds of databases on a single MetroCluster system. It's also simple to
use. The use of MetroCluster does not necessarily add to or change any
best practices for operating a enterprise applications and databases.
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.
```

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 Oracle RAC, Veritas Cluster Server (VCS) o VMware ha vengono spesso utilizzati per creare un cluster in tutti i siti, e in molti casi il processo di failover può essere guidato da semplici script.

In caso di perdita dei nodi primari, il clusterware (o gli script) viene configurato in modo da portare le applicazioni online nel sito alternativo. Un'opzione è creare server di standby preconfigurati per le risorse NFS o SAN che costituiscono l'applicazione. 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
4. Avvio dell'applicazione

Il requisito principale di questo approccio è rappresentato da un sistema operativo in esecuzione sul sito remoto. Deve essere preconfigurato con file binari delle applicazioni, il che significa anche che attività come l'applicazione di patch devono essere eseguite sul sito primario e di standby. In alternativa, è possibile eseguire il mirroring dei file binari dell'applicazione 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. 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. 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 Databases, 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.

Singola istanza di Oracle su MetroCluster

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 RAC esteso su MetroCluster

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 necessario un RAC esteso Active-Active, è necessario considerare il mirroring ASM al posto di MetroCluster. Il mirroring ASM 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 le LUN di avvio, compresi i dischi di avvio virtualizzati, vengono utilizzati con Oracle RAC, l' `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

Database Oracle con sincronizzazione attiva SnapMirror

SnapMirror Active Sync consente un RPO selettivo=mirroring sincrono di 0 KB per singoli database Oracle e ambienti applicativi.

SnapMirror Active Sync è essenzialmente una funzionalità SnapMirror migliorata per LA SAN che consente agli host di accedere a una LUN dal sistema che ospita il LUN e il sistema che ospita la sua replica.

SnapMirror Active Sync e SnapMirror Sync condividono un motore di replica, tuttavia SnapMirror Active Sync include funzionalità aggiuntive come il failover trasparente delle applicazioni e il failback per le applicazioni Enterprise.

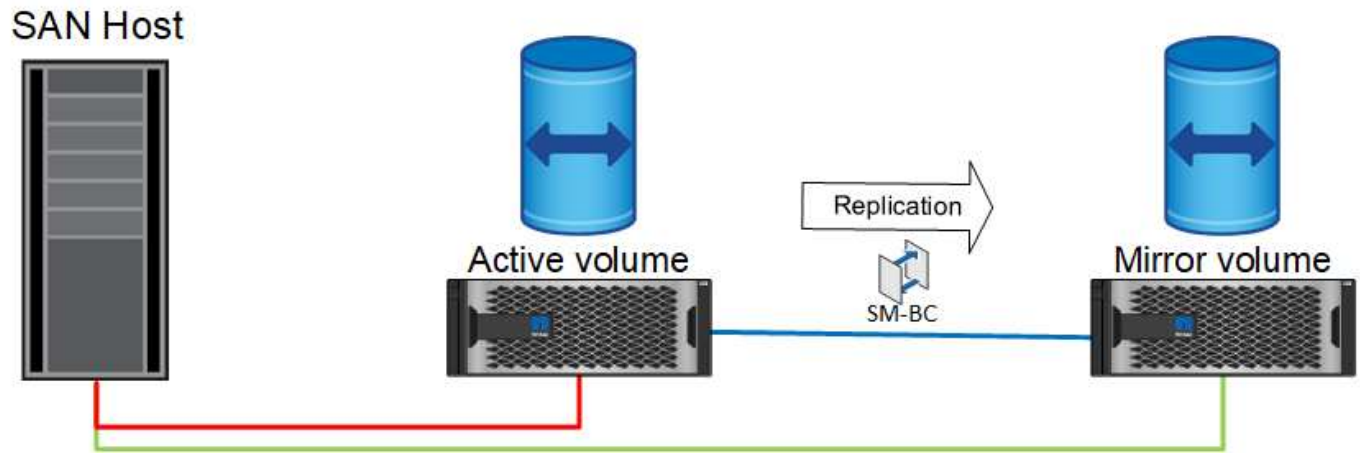
In pratica, funziona in modo simile a una versione granulare di MetroCluster, consentendo una replica sincrona RPO=0:1 selettiva e granulare per i singoli carichi di lavoro. Il comportamento del percorso di basso livello è molto diverso da MetroCluster, ma il risultato finale da un punto di vista dell'host è simile.

Accesso al percorso

Con SnapMirror Active Sync, i dispositivi di storage sono visibili per l'hosting dei sistemi operativi dagli array di storage primari e remoti. I percorsi vengono gestiti tramite l'ALUA (Asymmetric Logical Unit Access), un protocollo standard di settore per l'identificazione dei percorsi ottimizzati tra un sistema storage e un host.

Il percorso del dispositivo più breve per accedere all'i/o è considerato percorsi attivi/ottimizzati e il resto dei percorsi è considerato percorsi attivi/non ottimizzati.

La relazione di sincronizzazione attiva di SnapMirror è presente tra una coppia di SVM situate su cluster diversi. Entrambe le SVM sono in grado di fornire i dati, ma ALUA utilizza preferibilmente la SVM che attualmente è proprietaria dei dischi su cui risiedono le LUN. L'io alla SVM remota verrà fornito con un proxy attraverso l'interconnessione sincrona attiva di SnapMirror.



Replica sincrona

Durante le normali operazioni, la copia remota è una replica sincrona RPO=0/7, con un'unica eccezione. Se i dati non possono essere replicati, con la sincronizzazione attiva di SnapMirror libererà il requisito di replicare i dati e riprendere la fornitura io. Questa opzione è preferita dai clienti che considerano la perdita del collegamento di replica quasi un evento disastroso o che non desiderano arrestare le operazioni di business quando i dati non possono essere replicati.

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 scaricata dal supporto NetApp. Mediator automatizza le operazioni di failover sia per il cluster di storage del sito primario che per quello remoto. Può essere implementato su una piccola macchina virtuale (VM) ospitata on-premise o nel cloud. Una volta configurato, funge da terzo sito per monitorare gli scenari di failover per entrambi i siti.

Failover del database Oracle con SnapMirror Active Sync

Il motivo principale per ospitare un database Oracle su SnapMirror Active Sync è fornire il failover trasparente durante gli eventi di storage pianificati e non.

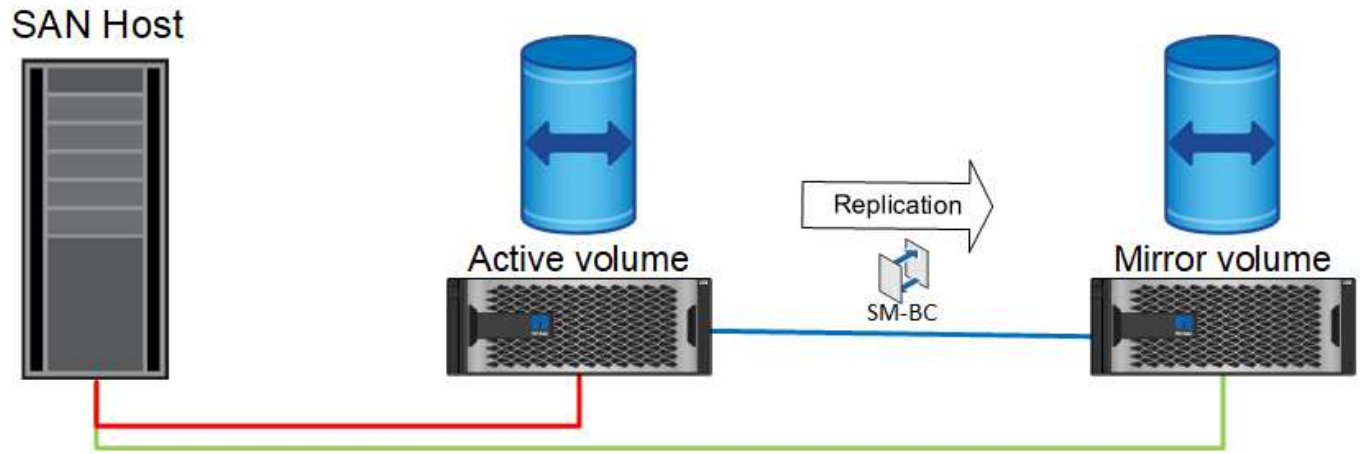
SnapMirror Active Sync supporta due tipi di operazioni di failover dello storage: Pianificate e meno, che funzionano in modi leggermente diversi. Un failover pianificato viene avviato manualmente dall'amministratore per uno switchover rapido verso un sito remoto, mentre il failover non pianificato viene avviato automaticamente dal mediatore del terzo sito. Lo scopo principale di un failover pianificato è quello di eseguire patch e aggiornamenti incrementali, eseguire test di disaster recovery o adottare una politica formale di commutazione delle operazioni tra i siti nel corso dell'anno per dimostrare la piena funzionalità di sincronizzazione attiva.

I diagrammi mostrano cosa accade durante le normali operazioni di failover e failback. Per maggiore facilità di illustrazione, sono raffigurati un LUN replicato. In una configurazione di sincronizzazione attiva di SnapMirror effettiva, la replica si basa sui volumi, dove ogni volume contiene una o più LUN, ma per semplificarne la visione, il livello del volume è stato rimosso.

Funzionamento normale

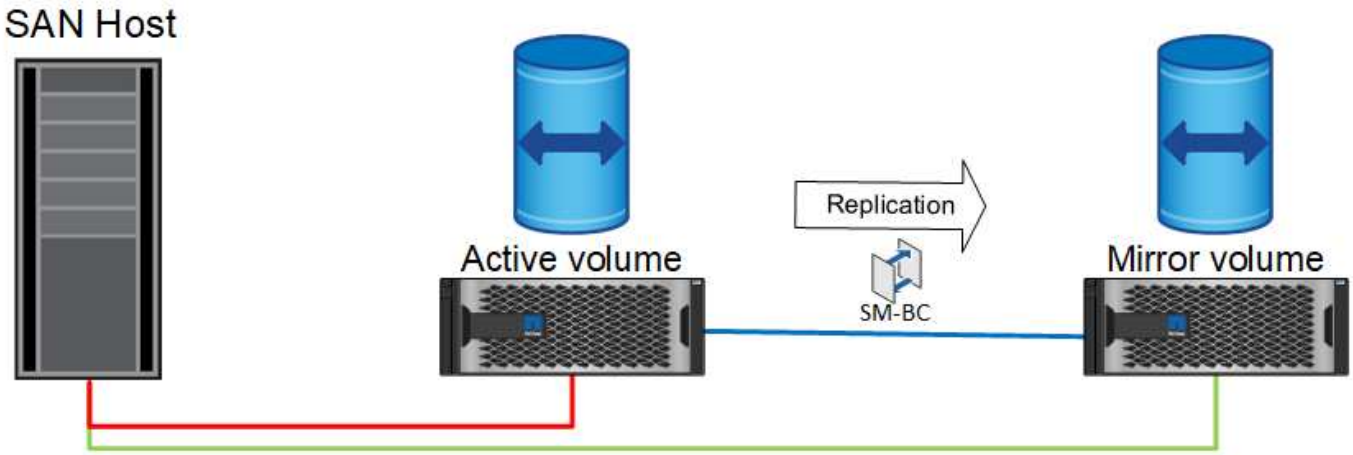
Durante il normale funzionamento, è possibile accedere a un LUN dalla replica locale o remota. La linea rossa indica il percorso ottimizzato come pubblicizzato da ALUA, e il risultato dovrebbe essere che io è preferenzialmente inviato lungo questo percorso.

La linea verde è un percorso attivo, ma richiede una maggiore latenza, perché i/o su quel percorso devono essere passati attraverso il percorso di sincronizzazione attivo di SnapMirror. La latenza aggiuntiva dipende dalla velocità dell'interconnessione tra i siti utilizzati per la sincronizzazione attiva di SnapMirror.



Guasto

Se la copia del mirror attivo non è più disponibile, a causa di un failover pianificato o non pianificato, ovviamente non sarà più utilizzabile. Tuttavia, il sistema remoto possiede una replica sincrona e i percorsi SAN verso il sito remoto esistono già. Il sistema remoto è in grado di gestire i/o per quel LUN.

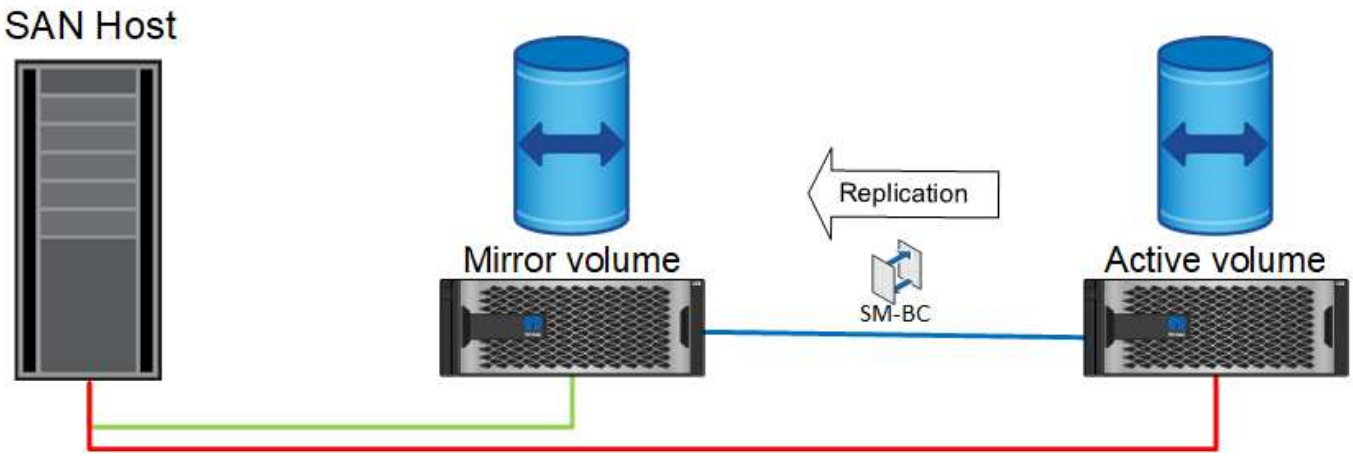


Failover

Il failover fa sì che la copia remota diventi la copia attiva. I percorsi vengono modificati da Active a Active/Optimized e l'io continua a essere gestito senza perdita di dati.

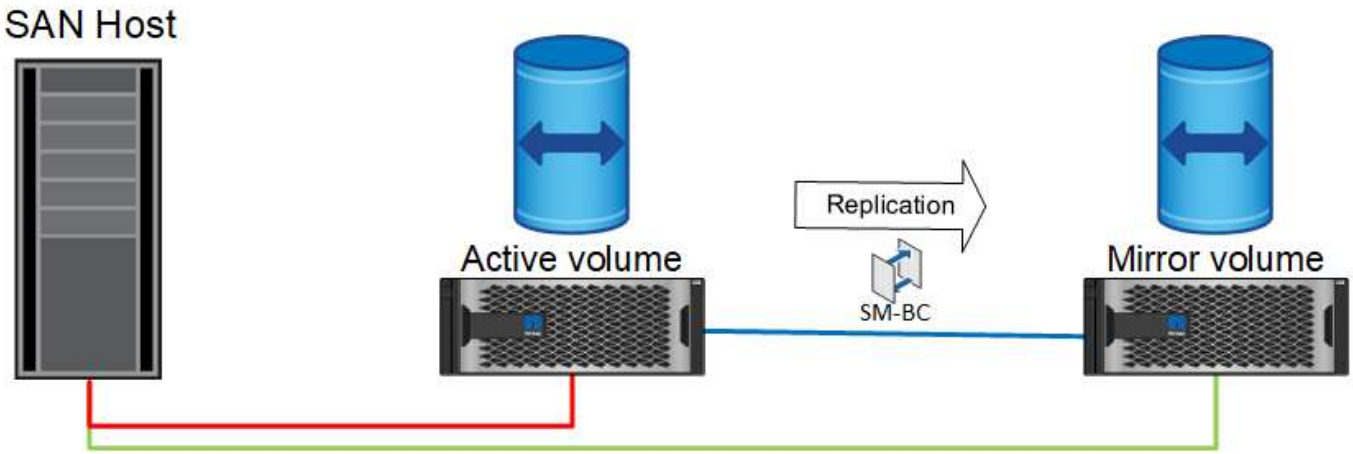
Riparare

Una volta che il sistema di origine è tornato in servizio, SnapMirror Active Sync può risincronizzare la replica, ma eseguendo l'altra direzione. Attualmente la configurazione è essenzialmente la stessa del punto di partenza, con la sola eccezione che i siti mirror attivi sono stati invertiti.



Failback

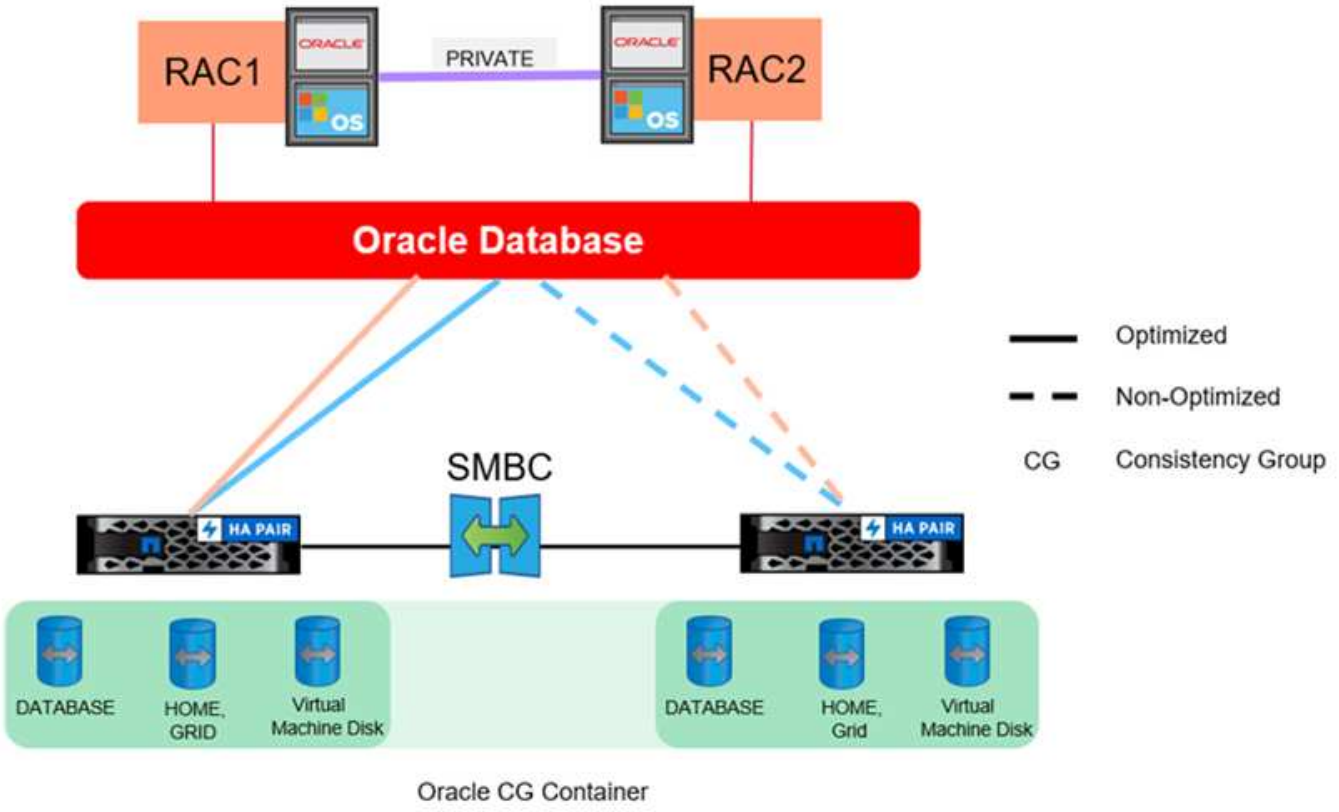
Se lo si desidera, un amministratore può eseguire un failback e riportare la copia attiva delle LUN nei controller originali.



Database Oracle a singola istanza con sincronizzazione attiva SnapMirror

Il diagramma seguente mostra un semplice modello di distribuzione in cui sono presenti dispositivi di storage con zoning o connessi dai cluster di storage primari e remoti per un database Oracle.

Oracle è configurato solo sul primario. Questo modello risolve il failover dello storage perfetto in caso di disastri sul lato dello storage, senza perdita di dati e senza downtime applicativi. Questo modello, tuttavia, non fornirebbe un'elevata disponibilità dell'ambiente di database durante un errore del sito. Questo tipo di architettura è utile per i clienti che cercano una soluzione senza perdita di dati con alta disponibilità dei servizi di storage, ma accettano che una perdita totale del cluster di database richieda lavoro manuale.



Questo approccio consente inoltre di risparmiare sui costi di licenza Oracle. La preconfigurazione dei nodi di database Oracle nel sito remoto richiede la licenza di tutti i core in base alla maggior parte dei contratti di licenza Oracle. Se il ritardo causato dal tempo richiesto per installare un server di database Oracle e montare la copia di dati rimanente è accettabile, questa progettazione può essere molto conveniente.

Oracle RAC con SnapMirror Active Sync

SnapMirror Active Sync offre un controllo granulare sulla replica del set di dati per scopi quali il bilanciamento del carico o il failover di una singola applicazione. L'architettura complessiva è simile a un cluster RAC esteso, ma alcuni database sono dedicati a siti specifici e il carico complessivo viene distribuito.

Ad esempio, puoi costruire un cluster Oracle RAC che ospita sei singoli database. Lo storage per tre dei database è principalmente ospitato sul sito A e quello per gli altri tre database sul sito B. Questa configurazione garantisce le migliori prestazioni possibili riducendo al minimo il traffico tra siti. Inoltre, le applicazioni vengono configurate in modo da utilizzare le istanze del database locali del sistema storage con percorsi attivi. In questo modo si riduce al minimo il traffico di interconnessione RAC. Infine, questa progettazione complessiva garantisce che tutte le risorse di calcolo vengano utilizzate in modo uniforme. Con il variare dei carichi di lavoro, è possibile eseguire selettivamente il failover dei database fra diversi siti, in modo da garantire un caricamento uniforme.

A parte la granularità, i principi e le opzioni di base per Oracle RAC che utilizzano la sincronizzazione attiva SnapMirror sono gli stessi di "Oracle RAC su MetroCluster"

Scenari di errori di sincronizzazione attiva per i database Oracle e SnapMirror

Esistono vari scenari di guasti di SnapMirror Active Sync (SM-AS), ciascuno con risultati diversi.

Scenario	Risultato
Errore del collegamento di replica	Mediatore riconosce questo scenario split-brain e riprende l'i/o sul nodo che contiene la copia master. Quando la connettività tra i siti è di nuovo online, il sito alternativo esegue la risincronizzazione automatica.
Guasto allo storage della sede principale	Il failover non pianificato automatizzato viene avviato da Mediator. Nessuna interruzione di i/O.
Errore dello storage nel sito remoto	Non si verifica alcuna interruzione di i/O. Si verifica una pausa momentanea a causa della rete che causa l'interruzione della replica di sincronizzazione e il master che stabilisce che è il legittimo proprietario continuare a servire i/o (consensus). Pertanto, si verifica una pausa i/o di alcuni secondi, quindi l'i/o riprenderà. Quando il sito è in linea, viene eseguita una risincronizzazione automatica.
Perdita di Mediator o collegamento tra Mediator e gli array di storage	L'i/o continua e rimane sincronizzato con il cluster remoto, ma in assenza di Mediator non è possibile eseguire il failover e il failback pianificati/non pianificati automatici.
Perdita di uno degli storage controller nel cluster ha	Il nodo partner nel cluster di ha tenta un takeover (NDO). Se il takeover ha esito negativo, Mediator nota che entrambi i nodi nello storage sono inattivi ed esegue un failover non pianificato automatico nel cluster remoto.
Perdita di dischi	L'io continua per un massimo di tre guasti consecutivi al disco. Questo fa parte di RAID-TEC.
Perdita dell'intero sito in un'implementazione tipica	I server sul sito in errore non saranno più disponibili. Le applicazioni che supportano il clustering possono essere configurate per l'esecuzione in entrambi i siti e la continuità delle operazioni sul sito alternativo, anche se la maggior parte di tali applicazioni richiede un tiebreaker a 3rd siti, in modo simile a quanto SM-AS richiede il mediatore. Senza cluster a livello di applicazione, le applicazioni dovranno essere avviate nel sito rimasto. Ciò influisce sulla disponibilità, ma viene mantenuto RPO=0. Non si perderebbero dati.

Informazioni sul copyright

Informazioni sul marchio commerciale