MetroCluster

Enterprise applications

NetApp
May 03, 2024

Tabla de contenidos

MetroCluster

Arquitectura física de MetroCluster y bases de datos Oracle

Para comprender el funcionamiento de las bases de datos Oracle en un entorno MetroCluster, es necesario explicar el diseño físico de un sistema MetroCluster.

Esta documentación sustituye al informe técnico *TR-4592 publicado anteriormente: Oracle en MetroCluster.*

MetroCluster está disponible en 3 configuraciones diferentes

- Pares DE ALTA DISPONIBILIDAD con conectividad IP
- Pares DE ALTA DISPONIBILIDAD con conectividad FC
- Controladora única con conectividad FC

[NOTA]El término 'conectividad' hace referencia a la conexión de cluster utilizada para la replicación entre sitios. No hace referencia a los protocolos de host. Todos los protocolos del lado del host se admiten como de costumbre en una configuración de MetroCluster, independientemente del tipo de conexión utilizada para la comunicación entre clústeres.

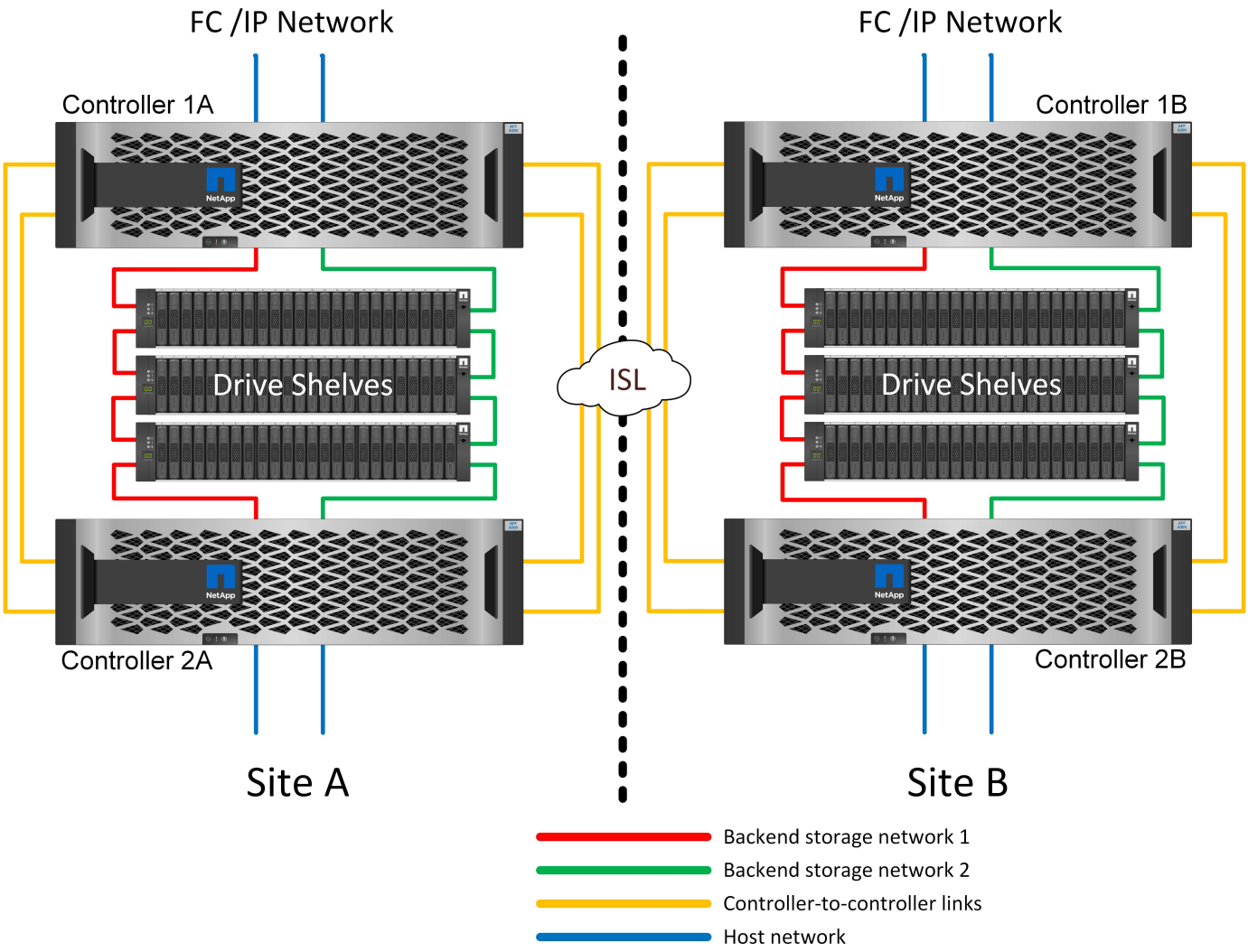
IP de MetroCluster

La configuración IP de MetroCluster para pares de alta disponibilidad utiliza dos o cuatro nodos por sitio. Esta opción de configuración aumenta la complejidad y los costes relacionados con la opción de dos nodos, pero ofrece una ventaja importante: La redundancia dentro del sitio. Un simple fallo de una controladora no requiere acceso a los datos a través de la WAN. El acceso a los datos sigue siendo local a través de la controladora local alternativa.

La mayoría de los clientes eligen la conectividad IP porque los requisitos de infraestructura son más simples. En el pasado, la conectividad entre sitios de alta velocidad solía ser más fácil de aprovisionar mediante switches FC y de fibra oscura; sin embargo, hoy en día los circuitos IP de alta velocidad y baja latencia son más fáciles de obtener.

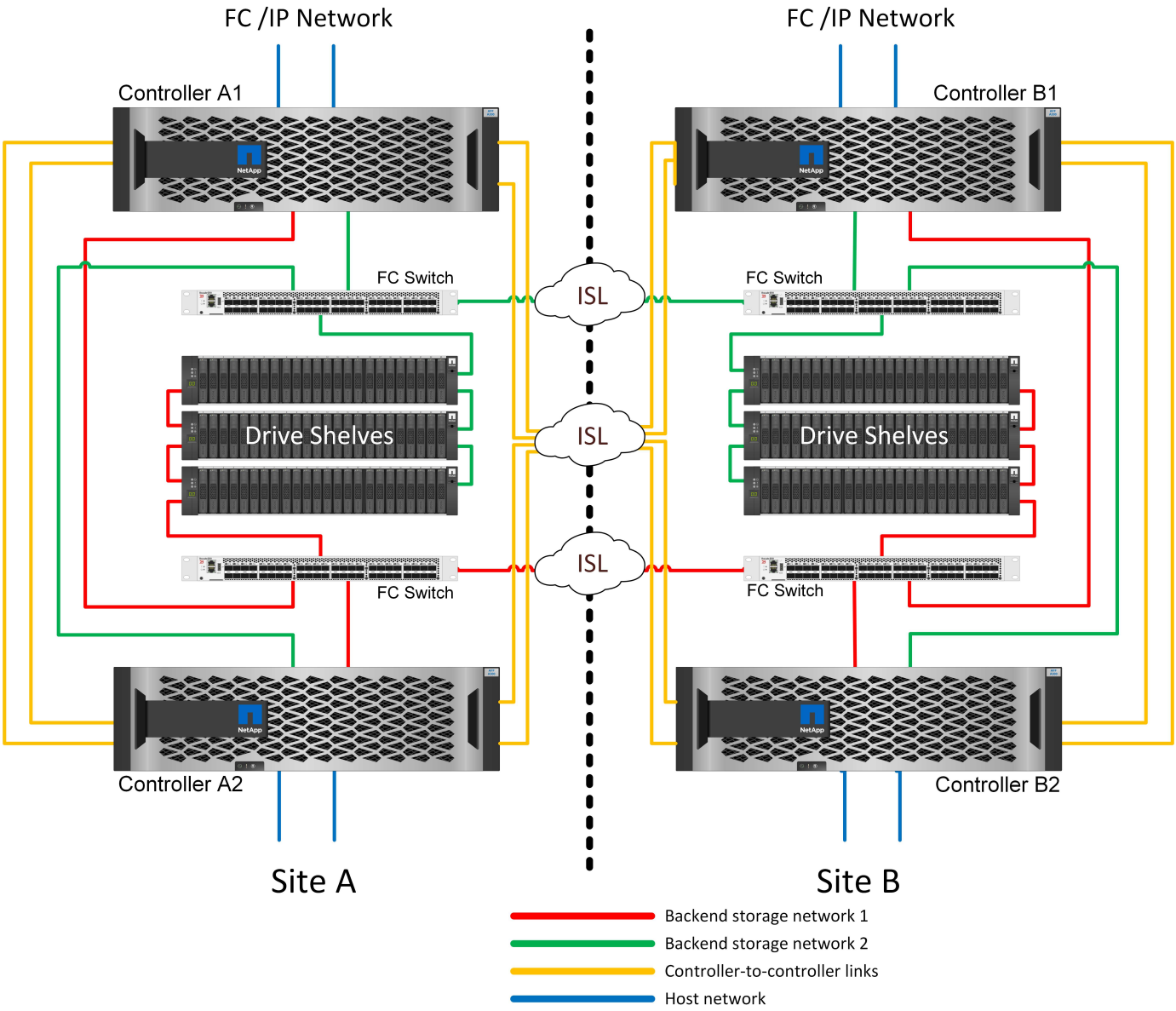
La arquitectura además es más sencilla ya que las únicas conexiones entre sitios son para las controladoras. En MetroCluster conectados a FC SAN, una controladora escribe directamente en las unidades del sitio opuesto y, por lo tanto, requiere conexiones SAN, switches y puentes adicionales. En cambio, una controladora con una configuración IP escribe en las unidades opuestas a través de la controladora.

Para obtener información adicional, consulte la documentación oficial de ONTAP y. "Arquitectura y diseño de la solución MetroCluster IP".



MetroCluster con conexión SAN FC de par de ALTA DISPONIBILIDAD

La configuración MetroCluster FC de par de alta disponibilidad utiliza dos o cuatro nodos por sitio. Esta opción de configuración aumenta la complejidad y los costes relacionados con la opción de dos nodos, pero ofrece una ventaja importante: La redundancia dentro del sitio. Un simple fallo de una controladora no requiere acceso a los datos a través de la WAN. El acceso a los datos sigue siendo local a través de la controladora local alternativa.



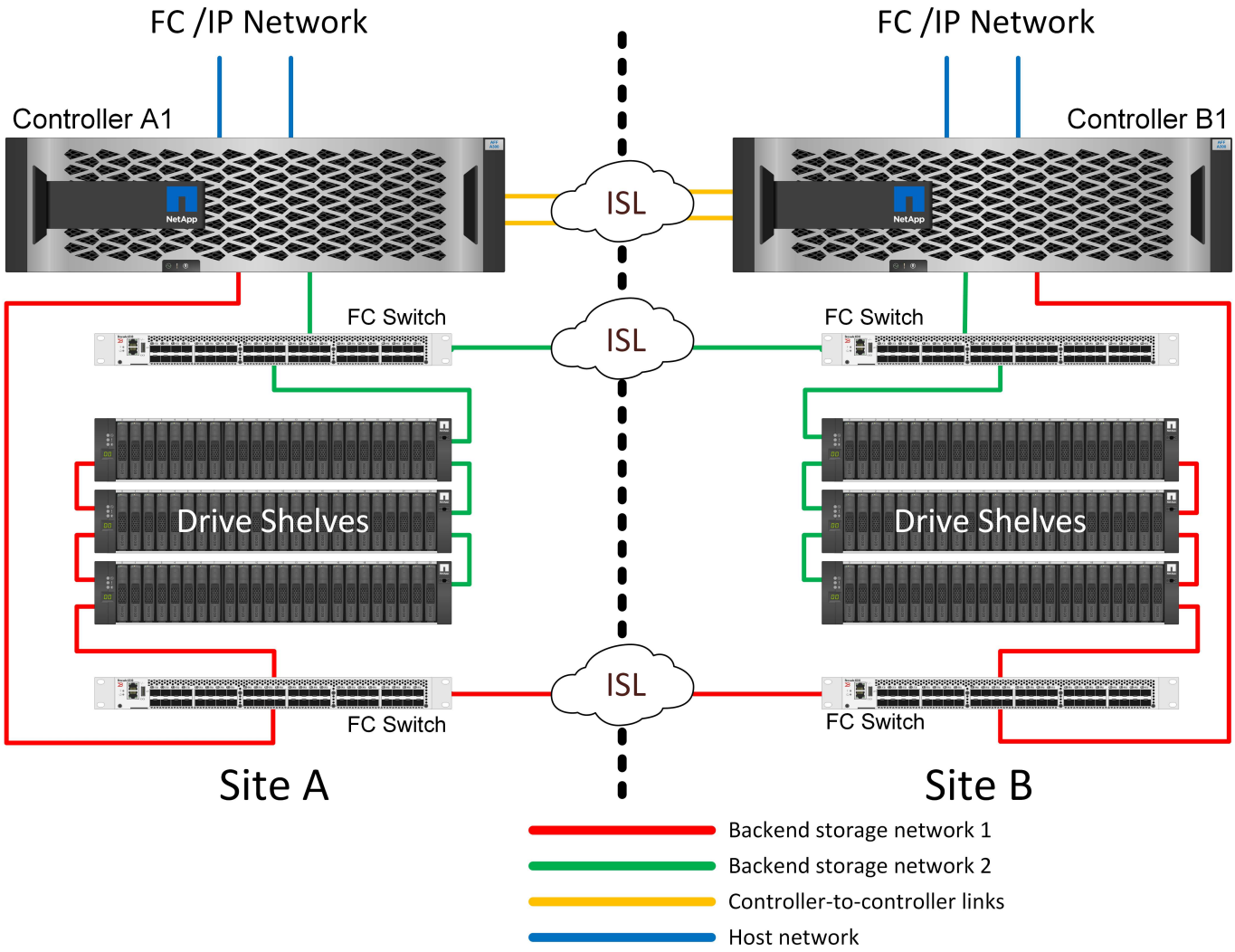
Algunas infraestructuras multisitio no están diseñadas para operaciones activo-activo, sino que se utilizan más como sitio principal y sitio de recuperación de desastres. En esta situación, generalmente es preferible una opción MetroCluster de una pareja de alta disponibilidad por las siguientes razones:

- Aunque un clúster MetroCluster de dos nodos es un sistema de alta disponibilidad, el fallo inesperado de una controladora o de tareas de mantenimiento planificadas requiere que los servicios de datos deban estar online en el sitio opuesto. Si la conectividad de red entre los sitios no puede soportar el ancho de banda requerido, el rendimiento se ve afectado. La única opción sería también conmutar por error los diversos sistemas operativos host y los servicios asociados a la ubicación alternativa. El clúster MetroCluster de la pareja de alta disponibilidad elimina este problema porque la pérdida de una controladora hace que la conmutación al respaldo sea sencilla dentro del mismo sitio.
- Algunas topologías de red no están diseñadas para el acceso entre sitios, sino que utilizan subredes diferentes o SAN FC aisladas. En estos casos, el clúster MetroCluster de dos nodos ya no funciona como un sistema de alta disponibilidad porque la controladora alternativa no puede proporcionar datos a los servidores del sitio opuesto. La opción MetroCluster de par de alta disponibilidad es necesaria para ofrecer una redundancia completa.
- Si se considera una infraestructura de dos sitios como una única infraestructura de alta disponibilidad, la configuración de MetroCluster de dos nodos es adecuada. Sin embargo, si el sistema debe funcionar

durante un largo período de tiempo tras el fallo del sitio, se prefiere un par de alta disponibilidad porque sigue proporcionando alta disponibilidad dentro de un único sitio.

MetroCluster FC de dos nodos conectado a SAN

La configuración de MetroCluster de dos nodos solo utiliza un nodo por sitio. Este diseño es más sencillo que la opción de pareja de alta disponibilidad porque hay menos componentes que configurar y mantener. También ha reducido las demandas de infraestructura en términos de cableado y conmutación FC. Por último, reduce los costes.



El impacto obvio de este diseño es que el fallo de una controladora en un único sitio significa que los datos están disponibles en el sitio opuesto. Esta restricción no es necesariamente un problema. Muchas empresas tienen operaciones de centros de datos multisitio con redes extendidas de alta velocidad y baja latencia que funcionan básicamente como una única infraestructura. En estos casos, la versión de dos nodos de MetroCluster es la configuración preferida. Varios proveedores de servicios utilizan actualmente sistemas de dos nodos a escala de petabytes.

Funcionalidades de resiliencia de MetroCluster

No hay puntos únicos de error en una solución de MetroCluster:

- Cada controladora tiene dos rutas independientes a las bandejas de unidades en el sitio local.

- Cada controladora tiene dos rutas independientes a las bandejas de unidades en el sitio remoto.
- Cada controladora tiene dos rutas independientes a las controladoras del sitio opuesto.
- En la configuración de par de alta disponibilidad, cada controladora tiene dos rutas desde su compañero local.

En resumen, puede eliminarse cualquier componente de la configuración sin poner en riesgo la capacidad de MetroCluster para suministrar datos. La única diferencia en términos de flexibilidad entre las dos opciones es que la versión del par de alta disponibilidad sigue siendo un sistema de almacenamiento de alta disponibilidad global tras un fallo del sitio.

Arquitectura lógica de MetroCluster y bases de datos de Oracle

Comprender el funcionamiento de las bases de datos Oracle en un entorno MetroCluster alsop requiere alguna explicación de la funcionalidad lógica de un sistema MetroCluster.

Protección contra errores del sitio: NVRAM y MetroCluster

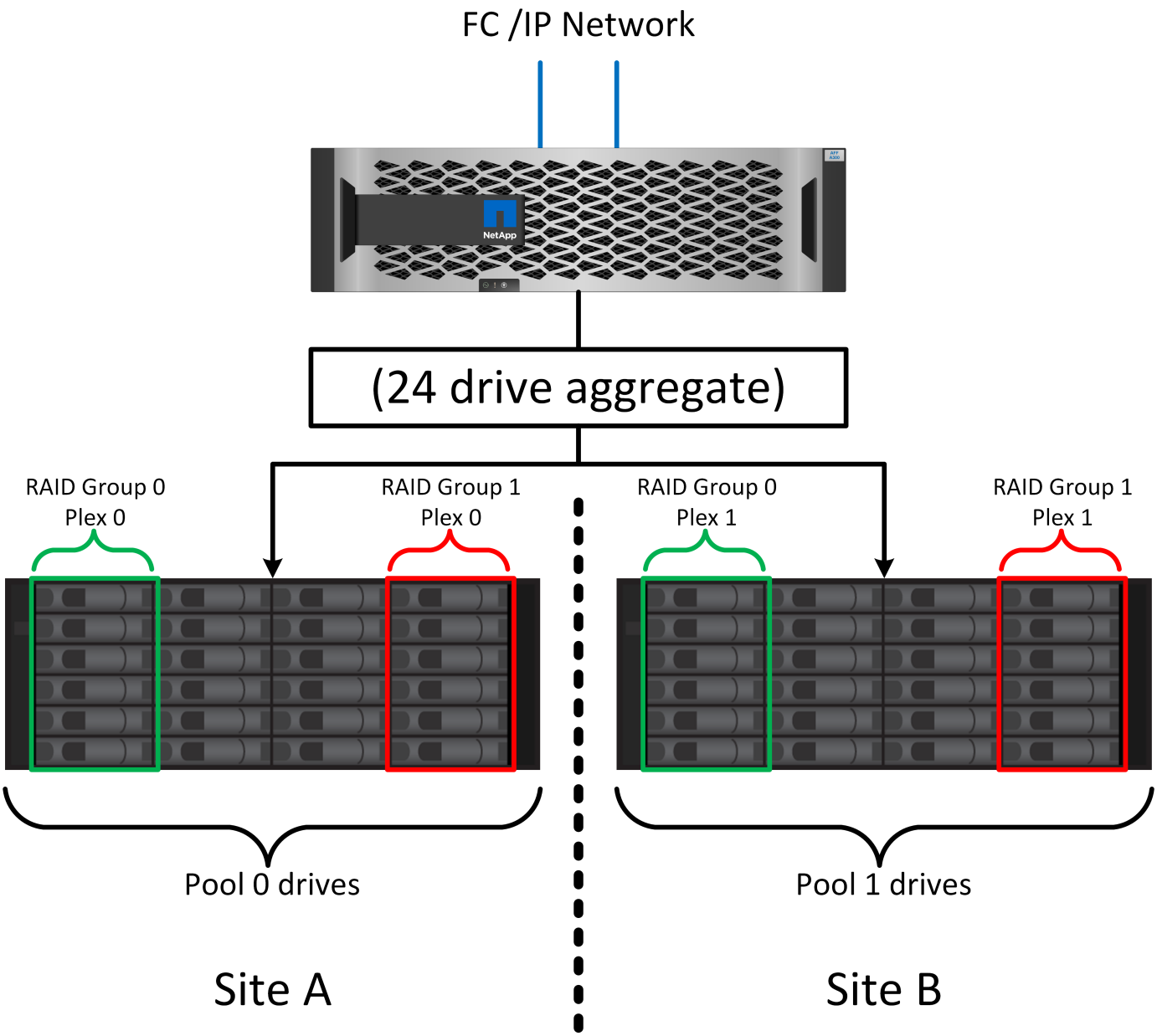
MetroCluster amplía la protección de datos de NVRAM de las siguientes formas:

- En una configuración de dos nodos, los datos de la NVRAM se replican mediante los enlaces Inter-Switch (ISL) al compañero remoto.
- En una configuración de par de alta disponibilidad, los datos de NVRAM se replican tanto en el partner local como en el remoto.
- La escritura no se reconoce hasta que se replica a todos los partners. Esta arquitectura protege la I/O en tránsito de fallos del sitio mediante la replicación de los datos de NVRAM en un partner remoto. Este proceso no está relacionado con la replicación de datos a nivel de unidad. La controladora propietaria de los agregados se encarga de la replicación de datos escribiendo en ambos complejos del agregado, pero seguirá habiendo protección contra la pérdida de I/O en tránsito en caso de pérdida del sitio. Los datos de NVRAM replicados solo se utilizan si una controladora asociada debe asumir el relevo de una controladora que ha fallado.

Protección frente a fallos de sitios y bandejas: SyncMirror y complejos

SyncMirror es una tecnología de mirroring que mejora, pero no sustituye, RAID DP ni RAID-TEC. Refleja el contenido de dos grupos RAID independientes. La configuración lógica es la siguiente:

1. Las unidades se configuran en dos pools según la ubicación. Un pool se compone de todas las unidades en el sitio A, y el segundo pool se compone de todas las unidades en el sitio B.
2. A continuación, se crea un pool de almacenamiento común, conocido como agregado, basado en conjuntos reflejados de grupos RAID. Se extrae un número igual de unidades en cada sitio. Por ejemplo, un agregado SyncMirror de 20 unidades estaría compuesto por 10 unidades del sitio A y 10 unidades del sitio B.
3. Cada conjunto de unidades en un sitio determinado se configura automáticamente como uno o varios grupos RAID DP o RAID-TEC completamente redundantes, independientemente del uso de mirroring. Este uso de RAID debajo del mirroring proporciona protección de datos incluso después de la pérdida de un sitio.



La figura anterior muestra una configuración de SyncMirror de ejemplo. Se creó un agregado de 24 unidades en la controladora con 12 unidades de una bandeja asignada en el sitio A y 12 unidades de una bandeja asignada en el sitio B. Las unidades se agruparon en dos grupos RAID reflejados. El grupo RAID 0 incluye un plex de 6 unidades en el sitio A reflejado en un plex de 6 unidades en el sitio B. Del mismo modo, el grupo RAID 1 incluye un plex de 6 unidades en el sitio A, duplicado en un plex de 6 unidades en el sitio B.

Normalmente, SyncMirror se utiliza para proporcionar mirroring remoto con sistemas MetroCluster, con una copia de los datos de cada sitio. En ocasiones, se ha utilizado para proporcionar un nivel adicional de redundancia en un único sistema. En particular, proporciona redundancia a nivel de bandeja. Una bandeja de unidades ya contiene fuentes de alimentación y controladoras duales y en general es poco más que chapa metálica, pero en algunos casos, la protección adicional puede estar garantizada. Por ejemplo, un cliente de NetApp ha puesto en marcha SyncMirror para una plataforma móvil de análisis en tiempo real que se usa durante las pruebas de automoción. El sistema se separó en dos racks físicos suministrados con fuentes de alimentación independientes y sistemas UPS independientes.

Fallo de redundancia: NVFAIL

Como hemos visto anteriormente, la escritura no se reconoce hasta que se haya iniciado sesión en la NVRAM local y NVRAM en al menos otra controladora. Este método garantiza que un fallo de hardware o una interrupción del suministro eléctrico no provoquen la pérdida de operaciones de I/O en tránsito Si la NVRAM local falla o la conectividad a otros nodos falla, los datos ya no se reflejarían.

Si la NVRAM local informa de un error, el nodo se apaga. Este apagado hace que se conmute al nodo de respaldo a la controladora asociada cuando se utilizan pares de alta disponibilidad. Con MetroCluster, el comportamiento depende de la configuración general elegida, pero puede dar lugar a una conmutación automática por error a la nota remota. En cualquier caso, no se pierden datos porque la controladora que experimenta el fallo no reconoció la operación de escritura.

Un fallo de conectividad entre sitios que bloquea la replicación de NVRAM en nodos remotos es una situación más complicada. Las escrituras ya no se replican en los nodos remotos y, de este modo, se crea la posibilidad de perder datos si se produce un error grave en una controladora. Lo que es más importante, si se intenta conmutar a un nodo diferente durante estas condiciones, se pierden datos.

El factor de control es si NVRAM está sincronizada. Si NVRAM está sincronizada, la conmutación al nodo de respaldo nodo a nodo se realizará de forma segura sin riesgo de pérdida de datos. En una configuración de MetroCluster, si la NVRAM y los complejos de agregado subyacentes están sincronizados, es seguro continuar con la conmutación de sitios sin riesgo de pérdida de datos.

ONTAP no permite una conmutación por error o una conmutación cuando los datos no están sincronizados a menos que se fuercen la conmutación por error o la conmutación. Al forzar un cambio en las condiciones de esta manera, se reconoce que los datos podrían dejarse atrás en la controladora original y que la pérdida de datos es aceptable.

Las bases de datos y otras aplicaciones son especialmente vulnerables a las corrupción si se fuerza una conmutación al respaldo o conmutación por error porque mantienen cachés internos más grandes de datos en el disco. Si se produce un failover forzado o un switchover forzado, los cambios previamente reconocidos se descartan efectivamente. El contenido de la cabina de almacenamiento retrocede efectivamente en el tiempo y el estado de la caché ya no refleja el estado de los datos del disco.

Para evitar esta situación, ONTAP permite configurar volúmenes para una protección especial contra un fallo de NVRAM. Cuando se activa, este mecanismo de protección hace que un volumen entre en un estado denominado NVFAIL. Este estado provoca errores de I/O que provocan un bloqueo de la aplicación. Este bloqueo hace que las aplicaciones se cierren para que no utilicen datos obsoletos. No se deben perder los datos porque los datos de transacción confirmados deben estar presentes en los registros. Los siguientes pasos habituales son para que un administrador apague completamente los hosts antes de volver a poner manualmente los LUN y los volúmenes de nuevo en línea. Aunque estos pasos pueden implicar cierto trabajo, este enfoque es la manera más segura de garantizar la integridad de los datos. No todos los datos requieren esta protección, por lo que el comportamiento NVFAIL se puede configurar volumen por volumen.

Pares DE ALTA disponibilidad y MetroCluster

MetroCluster está disponible en dos configuraciones: De dos nodos y de pareja de alta disponibilidad. La configuración de dos nodos se comporta igual que un par de alta disponibilidad con respecto a NVRAM. En caso de que falle repentinamente, el nodo asociado puede reproducir los datos de NVRAM para hacer que las unidades sean coherentes y asegurarse de que no se ha perdido ninguna escritura reconocida.

La configuración de par de alta disponibilidad replica la NVRAM también en el nodo del partner local. Un fallo de controladora sencillo provoca una reproducción de NVRAM en el nodo de partner, como es el caso con un par de alta disponibilidad independiente sin MetroCluster. En caso de pérdida repentina del sitio completo, el sitio remoto también cuenta con la NVRAM necesaria para hacer que las unidades sean coherentes y

empezar a servir datos.

Un aspecto importante de MetroCluster es que los nodos remotos no tienen acceso a los datos de los partners en condiciones operativas normales. Cada sitio funciona esencialmente como un sistema independiente que puede asumir la personalidad del sitio opuesto. Este proceso es conocido como una conmutación de sitios e incluye una conmutación de sitios planificada en la que las operaciones del sitio se migran de forma no disruptiva al sitio opuesto. También incluye situaciones no planificadas en las que se pierde un sitio y se requiere una conmutación por error manual o automática como parte de la recuperación ante desastres.

Conmutación de sitios y conmutación de estado

Los términos conmutación y conmutación de estado hacen referencia al proceso de transición de volúmenes entre controladoras remotas en una configuración de MetroCluster. Este proceso solo se aplica a los nodos remotos. Cuando MetroCluster se utiliza en una configuración de cuatro volúmenes, la conmutación por error de nodo local es el mismo proceso de toma de control y devolución descrito anteriormente.

Conmutación de sitios y conmutación de estado planificadas

Una conmutación de sitios o conmutación de estado planificada es similar a una toma de control o una conmutación al nodo primario entre nodos. El proceso tiene varios pasos y puede parecer que requiere varios minutos, pero lo que en realidad está sucediendo es una transición fluida multifase de los recursos de red y almacenamiento. El momento en que las transferencias de control se producen mucho más rápido que el tiempo necesario para que se ejecute el comando complete.

La principal diferencia entre toma de control/retorno al nodo primario y conmutación/conmutación de estado afecta a la conectividad SAN FC. Con la toma de control/devolución local, un host experimenta la pérdida de todas las rutas de FC hacia el nodo local y depende de su MPIO nativo para cambiar a las rutas alternativas disponibles. Los puertos no se reubican. Con la conmutación de sitios y la conmutación de estado, los puertos de destino FC virtuales en las controladoras se transfieren al otro sitio. De hecho, dejan de existir en la SAN durante un momento y luego vuelven a aparecer en una controladora alternativa.

Tiempo de espera de SyncMirror

SyncMirror es una tecnología de mirroring de ONTAP que proporciona protección contra fallos de bandeja. Cuando las bandejas se separan a lo largo de una distancia, el resultado es la protección de datos remota.

SyncMirror no ofrece mirroring síncrono universal. El resultado es una mejor disponibilidad. Algunos sistemas de almacenamiento utilizan mirroring constante todo o nada, llamado a veces modo domino. Esta forma de mirroring está limitada en la aplicación porque toda la actividad de escritura debe cesarse si se pierde la conexión con el sitio remoto. De lo contrario, una escritura existiría en un sitio, pero no en el otro. Normalmente, estos entornos están configurados para desconectar las LUN si se pierde la conectividad de sitio a sitio durante más de un breve período (como 30 segundos).

Este comportamiento es deseable para un pequeño subconjunto de entornos. Sin embargo, la mayoría de las aplicaciones requieren una solución que ofrezca replicación síncrona garantizada en condiciones de funcionamiento normales, pero con la posibilidad de suspender la replicación. Con frecuencia, se considera una pérdida total de conectividad entre sitios como una situación próxima a un desastre. Normalmente, estos entornos se mantienen online y proporcionan datos hasta que se repare la conectividad o se tome una decisión formal para desactivar el entorno para proteger los datos. Un requisito para el apagado automático de la aplicación solo debido a un fallo de replicación remota es inusual.

SyncMirror admite los requisitos de mirroring síncrono con la flexibilidad de un tiempo de espera agotado. Si se pierde la conectividad con el controlador remoto y/o plex, comienza la cuenta atrás con un temporizador de 30 segundos. Cuando el contador alcanza los 0, el procesamiento de I/O de escritura se reanuda utilizando los datos locales. La copia remota de los datos se puede utilizar, pero se congela en el tiempo hasta que se

restaure la conectividad. La resincronización aprovecha las copias Snapshot de nivel agregado para que el sistema vuelva al modo síncrono lo más rápido posible.

Cabe destacar que, en muchos casos, este tipo de replicación universal modo domino integral se implementa mejor en el nivel de aplicación. Por ejemplo, Oracle DataGuard incluye el modo de protección máxima, que garantiza la replicación de instancias largas en todas las circunstancias. Si el enlace de replicación falla durante un período que supera un tiempo de espera configurable, las bases de datos se cierran.

Cambio automático desatendido con Fabric Attached MetroCluster

La conmutación de sitios automática desatendida (AUSO) es una función MetroCluster conectada a estructuras que ofrece una forma de alta disponibilidad entre sitios. Como hemos visto anteriormente, MetroCluster está disponible en dos tipos: Una sola controladora en cada sitio o un par de alta disponibilidad en cada sitio. La principal ventaja de la opción de alta disponibilidad es que el apagado planificado o no planificado de la controladora sigue permitiendo que todas las operaciones de I/O sean locales. La ventaja de la opción de un único nodo es la reducción de los costes, la complejidad y la infraestructura.

El principal valor de AUSO es mejorar las funciones de alta disponibilidad de los sistemas MetroCluster Fabric Attached. Cada sitio monitorea el estado del sitio opuesto y, si no quedan nodos para servir datos, AUSO da como resultado un cambio rápido. Este método es especialmente útil en configuraciones de MetroCluster con solo un solo nodo por sitio porque acerca la configuración a un par de alta disponibilidad en términos de disponibilidad.

AUSO no puede ofrecer una supervisión completa a nivel de un par de alta disponibilidad. Un par de alta disponibilidad puede proporcionar una disponibilidad extremadamente alta porque incluye dos cables físicos redundantes para una comunicación directa entre nodos. Además, ambos nodos de un par de alta disponibilidad tienen acceso al mismo conjunto de discos en bucles redundantes, lo cual proporciona otra ruta para un nodo para supervisar el estado de otro.

Los clústeres de MetroCluster existen en todos los sitios en los que tanto la comunicación nodo a nodo como el acceso a disco dependen de la conectividad de red sitio a sitio. La capacidad de supervisar los latidos del resto del clúster es limitada. AUSO tiene que discriminar entre una situación en la que el otro sitio está realmente inactivo en lugar de no disponible debido a un problema de red.

Como resultado, una controladora de un par de alta disponibilidad puede emitir una toma de control si detecta un fallo de controladora que se produjo por un motivo específico, como un motivo de pánico en el sistema. También puede solicitar una toma de control si hay una pérdida completa de conectividad, a veces conocida como latido del corazón perdido.

Un sistema MetroCluster solo puede realizar de forma segura una conmutación automática cuando se detecta una falla específica en el sitio original. Además, la controladora que tome la propiedad del sistema de almacenamiento debe poder garantizar que los datos del disco y NVRAM estén sincronizados. El controlador no puede garantizar la seguridad de un cambio solo porque perdió el contacto con el sitio de origen, que podría estar operativo. Para ver opciones adicionales para automatizar una conmutación de sitios, consulte la información sobre la solución tiebreaker de MetroCluster (MCTB) en la siguiente sección.

Tiebreaker de MetroCluster con MetroCluster estructural

La "Tiebreaker de NetApp MetroCluster" El software puede ejecutarse en un tercer sitio para supervisar el estado del entorno de MetroCluster, enviar notificaciones y, opcionalmente, forzar una conmutación de sitios en caso de desastre. Puede encontrar una descripción completa del tiebreaker en la "Sitio de soporte de NetApp", Pero el propósito principal del MetroCluster tiebreaker es detectar la pérdida del sitio. También debe discriminar entre la pérdida del sitio y una pérdida de conectividad. Por ejemplo, la conmutación de sitios no debería ocurrir porque el tiebreaker no pudo llegar al sitio principal, por este motivo, tiebreaker también supervisa la capacidad del sitio remoto para comunicarse con el sitio principal.

El cambio automático con AUSO también es compatible con el MCTB. AUSO reacciona muy rápidamente porque está diseñado para detectar eventos de fallo específicos y luego invocar la conmutación de sitios solo cuando NVRAM y SyncMirror plexes están sincronizados.

Por el contrario, el desempate se encuentra de forma remota y, por lo tanto, debe esperar a que transcurra un temporizador antes de declarar un sitio muerto. El tiebreaker eventualmente detecta el tipo de fallo de la controladora cubierto por AUSO, pero en general AUSO ya ha iniciado la conmutación y posiblemente completado la conmutación antes de que actúe el tiebreaker. Se rechazaría el segundo comando de switchover resultante procedente del tiebreaker.

*Precaución: *El software MCTB no verifica que NVRAM estaba y/o los plexes estén sincronizados al forzar un cambio. La conmutación de sitios automática, si se configura, se debe deshabilitar durante actividades de mantenimiento que ocasionen la pérdida de sincronización para complejos de NVRAM o SyncMirror.

Además, es posible que el MCTB no solucione un desastre que lleve a la siguiente secuencia de eventos:

1. La conectividad entre sitios se interrumpe durante más de 30 segundos.
2. Se agota el tiempo de espera de la replicación de SyncMirror y las operaciones continúan en el sitio principal, dejando la réplica remota obsoleta.
3. Se pierde el sitio principal.El resultado es la presencia de cambios no replicados en el sitio principal. Una conmutación de sitios puede ser indeseable por varios motivos, entre los que se incluyen los siguientes:
    - Pueden haber datos cruciales en el sitio principal y esos datos podrían ser recuperables en algún momento. Un cambio que permitiera a la aplicación seguir funcionando descartaría esos datos cruciales.
    - Una aplicación del sitio superviviente que utilizaba recursos de almacenamiento en el sitio principal en el momento de la pérdida del sitio podría haber almacenado datos en caché. Un switchover introduciría una versión obsoleta de los datos que no coincide con la caché.
    - Un sistema operativo del sitio superviviente que utilizaba recursos de almacenamiento en el sitio principal en el momento de la pérdida del sitio podría haber almacenado los datos en caché. Un switchover introduciría una versión obsoleta de los datos que no coincide con la caché. La opción más segura es configurar el tiebreaker para que envíe una alerta si detecta un fallo del sitio y luego hacer que una persona tome una decisión sobre si forzar un cambio. Es posible que las aplicaciones o los sistemas operativos deban apagarse primero para borrar cualquier dato almacenado en caché. Además, la configuración NVFAIL puede usarse para agregar más protección y ayudar a simplificar el proceso de conmutación por error.

Mediador ONTAP con MetroCluster IP

El Mediador ONTAP se utiliza con MetroCluster IP y otras soluciones ONTAP. Funciona como un servicio tradicional de tiebreaker, muy similar al software MetroCluster tiebreaker de referencia anteriormente, pero también incluye una característica crítica, con la posibilidad de realizar una conmutación de sitios automatizada sin supervisión.

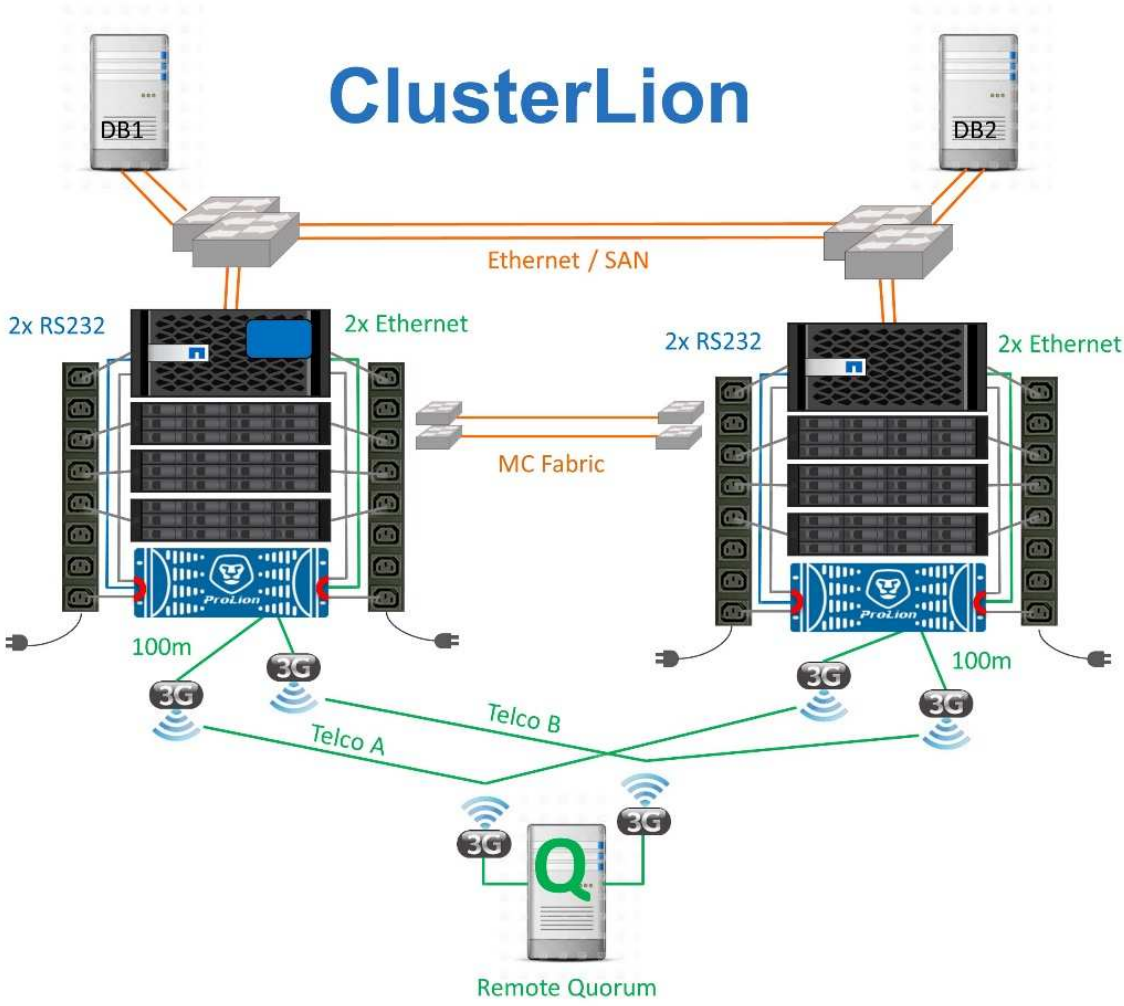
Una MetroCluster conectada a estructura tiene acceso directo a dispositivos de almacenamiento en el sitio opuesto. Esto permite que una controladora MetroCluster supervise el estado de las otras controladoras mediante la lectura de datos de latidos de las unidades. Esto permite que una controladora reconozca el fallo de otra controladora y realizar una conmutación por error.

Por el contrario, la arquitectura IP de MetroCluster enruta todas las I/O de forma exclusiva a través de la conexión del controlador; no hay acceso directo a los dispositivos de almacenamiento en el sitio remoto. Esto limita la capacidad de un controlador para detectar fallos y realizar una conmutación de sitios. Por lo tanto, el Mediador de ONTAP es necesario como dispositivo tiebreaker para detectar la pérdida del sitio y realizar

automáticamente una conmutación.

Tercer sitio virtual con ClusterLion

ClusterLion es un dispositivo de supervisión MetroCluster avanzado que funciona como un tercer sitio virtual. Este enfoque permite implementar MetroCluster de forma segura en una configuración de dos sitios con capacidad de conmutación de sitios totalmente automatizada. Además, ClusterLion puede realizar una supervisión de nivel de red adicional y ejecutar operaciones posteriores a la conmutación. La documentación completa está disponible en ProLion.



- Los dispositivos ClusterLion supervisan el estado de las controladoras con cables Ethernet y serie conectados directamente.
- Los dos aparatos están conectados entre sí con conexiones inalámbricas redundantes de 3G.
- La alimentación al controlador ONTAP se dirige a través de relés internos. En caso de un fallo del sitio, ClusterLion, que contiene un sistema UPS interno, corta las conexiones de alimentación antes de invocar un cambio. Este proceso garantiza que no se produzca ninguna condición cerebral dividida.
- ClusterLion realiza un switchover dentro del tiempo de espera de SyncMirror de 30 segundos o no lo hace en absoluto.
- ClusterLion no realiza una conmutación de sitios a menos que los estados de NVRAM y los complejos SyncMirror estén sincronizados.
- Dado que ClusterLion solo realiza una operación de switchover si MetroCluster está totalmente

sincronizado, no es necesario NVFAIL. Esta configuración permite que los entornos de expansión de sitios, como un Oracle RAC ampliado, permanezcan en línea, incluso durante una conmutación de sitios no planificada.

- El soporte incluye MetroCluster FAS e MetroCluster IP

Bases de datos de Oracle con SyncMirror

La base de la protección de datos de Oracle con un sistema MetroCluster es SyncMirror, una tecnología de mirroring síncrono de escalado horizontal y máximo rendimiento.

Protección de datos con SyncMirror

En el nivel más sencillo, la replicación síncrona implica que se debe realizar cualquier cambio en ambas partes del almacenamiento reflejado antes de que se reconozca. Por ejemplo, si una base de datos está escribiendo un registro o se está aplicando la revisión a un invitado VMware, no se debe perder nunca una escritura. Como nivel de protocolo, el sistema de almacenamiento no debe reconocer la escritura hasta que se haya comprometido a medios no volátiles en ambos sitios. Solo entonces es seguro proceder sin el riesgo de pérdida de datos.

El uso de una tecnología de replicación síncrona es el primer paso para diseñar y gestionar una solución de replicación síncrona. Lo más importante es comprender qué podría suceder durante varios escenarios de fallos planificados y no planificados. No todas las soluciones de replicación síncrona ofrecen las mismas funcionalidades. Si necesita una solución que proporcione un objetivo de punto de recuperación (RPO) de cero, lo que significa cero pérdida de datos, deben tenerse en cuenta todos los escenarios de fallo. En particular, ¿cuál es el resultado esperado cuando la replicación es imposible debido a la pérdida de conectividad entre sitios?

Disponibilidad de datos SyncMirror

La replicación de MetroCluster se basa en la tecnología de NetApp SyncMirror, que se ha diseñado para alternar eficientemente entre el modo síncrono y este se sale de él. Esta funcionalidad satisface los requisitos de los clientes que demandan replicación síncrona pero que también necesitan una alta disponibilidad para sus servicios de datos. Por ejemplo, si la conectividad con un sitio remoto se interrumpe, generalmente es preferible que el sistema de almacenamiento siga funcionando en un estado sin replicar.

Muchas soluciones de replicación síncrona solo pueden funcionar en modo síncrono. Este tipo de replicación compuesta por todos o nada se denomina a veces modo domino. Este tipo de sistemas de almacenamiento dejan de servir datos en lugar de permitir que las copias locales y remotas de datos se dessincronicen. Si la replicación se interrumpe de forma forzada, la resincronización puede requerir mucho tiempo y puede dejar al cliente expuesto a la pérdida de datos durante el tiempo que se restablece el mirroring.

SyncMirror no solo puede salir del modo síncrono sin problemas si no se puede acceder al sitio remoto, sino que también puede volver a sincronizar rápidamente con un estado RPO = 0 cuando se restaura la conectividad. La copia obsoleta de los datos en el sitio remoto también se puede conservar en estado utilizable durante la resincronización, lo que garantiza la existencia de copias locales y remotas de los datos en todo momento.

Cuando se requiere el modo domino, NetApp ofrece SnapMirror síncrono (SM-S). También existen opciones de nivel de aplicación, como Oracle DataGuard o timeouts ampliados para el mirroring de discos del host. Consulte con su equipo de cuentas de partner o de NetApp para obtener más información y opciones.

Conmutación al nodo de respaldo de bases de datos de 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.
Siguen siendo aplicables las prácticas recomendadas habituales y, si sus
necesidades solo requieren protección de datos con objetivo de punto de
recuperación = 0, esta se cumplirá con MetroCluster. Sin embargo, la
mayoría de los clientes utilizan MetroCluster no solo para la protección
de datos con objetivo de punto de recuperación = 0, sino también para
mejorar el objetivo de tiempo de recuperación durante escenarios de
desastre, y proporcionar una conmutación por error transparente como parte
de las actividades de mantenimiento del sitio.
```

Conmutación al nodo de respaldo con un SO preconfigurado

SyncMirror ofrece una copia síncrona de los datos del sitio de recuperación de desastres, pero para que los datos estén disponibles, requiere un sistema operativo y las aplicaciones asociadas. La automatización básica puede mejorar drásticamente el tiempo de conmutación al nodo de respaldo del entorno global. Los productos de Clusterware como Oracle RAC, Veritas Cluster Server (VCS) o VMware HA se utilizan a menudo para crear un clúster en todos los sitios y, en muchos casos, el proceso de conmutación por error se puede realizar con scripts sencillos.

Si se pierden los nodos primarios, el clusterware (o scripts) se configura para poner las aplicaciones en línea en el sitio alternativo. Una opción es crear servidores en espera que estén preconfigurados para los recursos NFS o SAN que componen la aplicación. Si el sitio principal falla, el clusterware o la alternativa con secuencia de comandos realiza una secuencia de acciones similar a las siguientes:

1. Forzar un cambio de MetroCluster
2. Detección de LUN FC (solo SAN)
3. Montaje de sistemas de archivos
4. Iniciando la aplicación

El requisito principal de este método es un sistema operativo en ejecución instalado en el sitio remoto. Se debe preconfigurar con binarios de aplicación, lo que también significa que las tareas como la aplicación de parches se deben realizar en la ubicación primaria y en espera. Como alternativa, los archivos binarios de la aplicación pueden duplicarse en el sitio remoto y montarse si se declara un desastre.

El procedimiento de activación real es simple. Los comandos como la detección de LUN sólo requieren unos pocos comandos por puerto FC. El montaje del sistema de archivos no es más que un `mount` Y tanto las bases de datos como ASM se pueden iniciar y parar en la CLI con un único comando. Si los volúmenes y los sistemas de archivos no se están utilizando en el sitio de recuperación de desastres antes de la conmutación de sitios, no es necesario establecerlos `dr-force- nvfail` en los volúmenes.

Conmutación por error con un sistema operativo virtualizado

La conmutación por error de los entornos de base de datos puede ampliarse para incluir el propio sistema operativo. En teoría, esta recuperación tras fallos se puede realizar con las LUN de arranque, pero la mayoría de las veces se realiza con un sistema operativo virtualizado. El procedimiento es similar a los siguientes pasos:

1. Forzar un cambio de MetroCluster
2. Montar los almacenes de datos que alojan las máquinas virtuales del servidor de bases de datos
3. Inicio de las máquinas virtuales
4. Iniciar las bases de datos manualmente o configurar las máquinas virtuales para iniciar automáticamente las bases de datos

Por ejemplo, un clúster ESX podría abarcar sitios. En caso de desastre, los equipos virtuales pueden conectarse en línea en el sitio de recuperación ante desastres después del cambio. Mientras los almacenes de datos que alojan los servidores de bases de datos virtualizadas no estén en uso en el momento del desastre, no es necesario configurarlos `dr-force- nvfail` en los volúmenes asociados.

Bases de datos de Oracle, MetroCluster y NVFAIL

NVFAIL es una función general de integridad de los datos en ONTAP que se ha diseñado para maximizar la protección de la integridad de los datos con las bases de datos.

En esta sección se amplía la explicación del NVFAIL básico de ONTAP para tratar temas específicos de MetroCluster.

Con MetroCluster, no se reconoce la escritura hasta que se haya iniciado sesión en la NVRAM y NVRAM locales en al menos otra controladora. Este método garantiza que un fallo de hardware o una interrupción del suministro eléctrico no provoquen la pérdida de operaciones de I/O en tránsito Si la NVRAM local falla o la conectividad a otros nodos falla, los datos ya no se reflejarían.

Si la NVRAM local informa de un error, el nodo se apaga. Este apagado hace que se conmute al nodo de respaldo a la controladora asociada cuando se utilizan pares de alta disponibilidad. Con MetroCluster, el comportamiento depende de la configuración general elegida, pero puede dar lugar a una conmutación automática por error a la nota remota. En cualquier caso, no se pierden datos porque la controladora que experimenta el fallo no reconoció la operación de escritura.

Un fallo de conectividad entre sitios que bloquea la replicación de NVRAM en nodos remotos es una situación más complicada. Las escrituras ya no se replican en los nodos remotos y, de este modo, se crea la posibilidad de perder datos si se produce un error grave en una controladora. Lo que es más importante, si se intenta conmutar a un nodo diferente durante estas condiciones, se pierden datos.

El factor de control es si NVRAM está sincronizada. Si NVRAM está sincronizada, la conmutación al nodo de respaldo nodo a nodo se realizará de forma segura sin riesgo de pérdida de datos. En una configuración de MetroCluster, si la NVRAM y los complejos de agregado subyacentes están sincronizados, es seguro continuar con la conmutación sin el riesgo de perder los datos.

ONTAP no permite una conmutación por error o una conmutación cuando los datos no están sincronizados a menos que se fuercen la conmutación por error o la conmutación. Al forzar un cambio en las condiciones de esta manera, se reconoce que los datos podrían dejarse atrás en la controladora original y que la pérdida de datos es aceptable.

Las bases de datos son especialmente vulnerables a los daños si se fuerza una conmutación por error o una conmutación por error porque las bases de datos mantienen cachés internos mayores de los datos en el disco. Si se produce un failover forzado o un switchover forzado, los cambios previamente reconocidos se descartan efectivamente. El contenido de la cabina de almacenamiento retrocede efectivamente en el tiempo y el estado de la caché de base de datos ya no refleja el estado de los datos del disco.

Para proteger aplicaciones contra esta situación, ONTAP permite configurar volúmenes para obtener protección especial contra un fallo NVRAM. Cuando se activa, este mecanismo de protección hace que un volumen entre en un estado denominado NVFAIL. Este estado provoca errores de I/O que provocan el cierre de la aplicación para que no utilicen datos obsoletos. No se deben perder los datos, ya que aún hay escrituras reconocidas en el sistema de almacenamiento y, con bases de datos, todos los datos de transacciones confirmados deben estar presentes en los registros.

Los siguientes pasos habituales son para que un administrador apague completamente los hosts antes de volver a poner manualmente los LUN y los volúmenes de nuevo en línea. Aunque estos pasos pueden implicar cierto trabajo, este enfoque es la manera más segura de garantizar la integridad de los datos. No todos los datos requieren esta protección, por lo que el comportamiento NVFAIL se puede configurar volumen por volumen.

NVFAIL forzado manualmente

La opción más segura para forzar una conmutación por error con un clúster de aplicaciones (incluido VMware, Oracle RAC y otros) que se distribuye entre los sitios es especificar `-force-nvfail-all` en la línea de comandos. Esta opción está disponible como medida de emergencia para garantizar que todos los datos almacenados en caché están vaciados. Si un host utiliza recursos de almacenamiento ubicados originalmente en el sitio afectado por desastres, recibirá errores de I/O o un identificador de archivos obsoleto (`ESTALE`) error. Las bases de datos de Oracle se bloquean y los sistemas de archivos se desconectan por completo o cambian al modo de sólo lectura.

Una vez finalizada la operación de switchover, el `in-nvfailed-state` La marca debe borrarse y las LUN deben colocarse en línea. Una vez finalizada esta actividad, se puede reiniciar la base de datos. Estas tareas se pueden automatizar para reducir el RTO.

dr-force-nvfail

Como medida de seguridad general, configure el `dr-force-nvfail` marque todos los volúmenes a los que se pueda acceder desde un sitio remoto durante las operaciones normales, lo que significa que se deben usar antes de la conmutación al respaldo. El resultado de esta configuración es que la selección de volúmenes remotos deja de estar disponible cuando se introducen `in-nvfailed-state` durante una conmutación de sitios. Una vez finalizada la operación de switchover, el `in-nvfailed-state` La marca debe borrarse y las LUN deben colocarse en línea. Una vez finalizadas estas actividades, se pueden reiniciar las aplicaciones. Estas tareas se pueden automatizar para reducir el RTO.

El resultado es como usar el `-force-nvfail-all` indicador para conmutadores manuales. Sin embargo, la cantidad de volúmenes afectados puede limitarse a solo los volúmenes que deben protegerse de aplicaciones o sistemas operativos que tienen caché anticuada.

Hay dos requisitos críticos para un entorno que no utiliza `dr-force-nvfail` en volúmenes de aplicaciones:

- Una conmutación de sitios forzada no debe ocurrir más de 30 segundos después de la pérdida del sitio principal.
- Una conmutación de sitios no debe producirse durante las tareas de mantenimiento ni ninguna otra condición en la que los plexes de SyncMirror o la replicación de NVRAM no estén sincronizados. El primer

requisito se puede cumplir con el uso de un software tiebreaker configurado para realizar una conmutación de sitios en un plazo de 30 segundos tras un fallo del sitio. Este requisito no significa que el cambio deba realizarse dentro de los 30 segundos posteriores a la detección de un fallo del centro. Esto significa que ya no es seguro forzar un cambio si han transcurrido 30 segundos desde que se confirmó que un sitio está operativo.

El segundo requisito se puede cumplir parcialmente deshabilitando todas las funcionalidades de conmutación automática de sitios cuando se sabe que la configuración de MetroCluster está fuera de sincronización. Mejor opción sería tener una solución tiebreaker que pueda supervisar el estado de la replicación de NVRAM y los plexes de SyncMirror. Si el clúster no está completamente sincronizado, tiebreaker no debería activar una conmutación de sitios.

El software NetApp MCTB no puede supervisar el estado de sincronización, por lo que debe desactivarse cuando MetroCluster no está sincronizado por cualquier motivo. ClusterLion incluye funcionalidades de supervisión de NVRAM y supervisión plex, y se puede configurar para no activar la conmutación de sitios a menos que se haya confirmado que el sistema MetroCluster está totalmente sincronizado.

Instancia única de Oracle en MetroCluster

Como se indicó anteriormente, la presencia de un sistema MetroCluster no necesariamente agrega ni cambia ninguna práctica recomendada para el funcionamiento de una base de datos. La mayoría de las bases de datos que se ejecutan actualmente en los sistemas MetroCluster del cliente son de única instancia y sigue las recomendaciones de la documentación de Oracle en ONTAP.

Conmutación al nodo de respaldo con un SO preconfigurado

SyncMirror ofrece una copia síncrona de los datos del sitio de recuperación de desastres, pero para que los datos estén disponibles, requiere un sistema operativo y las aplicaciones asociadas. La automatización básica puede mejorar drásticamente el tiempo de conmutación al nodo de respaldo del entorno global. Los productos de Clusterware, como Veritas Cluster Server (VCS), se utilizan a menudo para crear un clúster en todos los sitios y, en muchos casos, el proceso de conmutación por error se puede llevar a cabo con scripts sencillos.

Si se pierden los nodos primarios, el clusterware (o scripts) se configura para poner las bases de datos en línea en el sitio alternativo. Una opción es crear servidores en espera que estén preconfigurados para los recursos NFS o SAN que componen la base de datos. Si el sitio principal falla, el clusterware o la alternativa con secuencia de comandos realiza una secuencia de acciones similar a las siguientes:

1. Forzar un cambio de MetroCluster
2. Detección de LUN FC (solo SAN)
3. Montaje de sistemas de archivos y/o montaje de grupos de discos ASM
4. Iniciando la base de datos

El requisito principal de este método es un sistema operativo en ejecución instalado en el sitio remoto. Se debe preconfigurar con binarios de Oracle, lo que también significa que las tareas como los parches de Oracle se deben realizar en la ubicación primaria y en espera. Como alternativa, los binarios de Oracle se pueden duplicar en la ubicación remota y montar si se declara un desastre.

El procedimiento de activación real es simple. Los comandos como la detección de LUN sólo requieren unos pocos comandos por puerto FC. El montaje del sistema de archivos no es más que un `mount` Y tanto las bases de datos como ASM se pueden iniciar y parar en la CLI con un único comando. Si los volúmenes y los

sistemas de archivos no se están utilizando en el sitio de recuperación de desastres antes de la conmutación de sitios, no es necesario establecerlos `dr-force- nvfail` en los volúmenes.

Conmutación por error con un sistema operativo virtualizado

La conmutación por error de los entornos de base de datos puede ampliarse para incluir el propio sistema operativo. En teoría, esta recuperación tras fallos se puede realizar con las LUN de arranque, pero la mayoría de las veces se realiza con un sistema operativo virtualizado. El procedimiento es similar a los siguientes pasos:

1. Forzar un cambio de MetroCluster
2. Montar los almacenes de datos que alojan las máquinas virtuales del servidor de bases de datos
3. Inicio de las máquinas virtuales
4. Iniciar bases de datos manualmente o configurar las máquinas virtuales para iniciar automáticamente las bases de datos, por ejemplo, un clúster ESX puede abarcar varios sitios. En caso de desastre, los equipos virtuales pueden conectarse en línea en el sitio de recuperación ante desastres después del cambio. Mientras los almacenes de datos que alojan los servidores de bases de datos virtualizadas no estén en uso en el momento del desastre, no es necesario configurarlos `dr-force- nvfail` en los volúmenes asociados.

Oracle RAC ampliado en MetroCluster

Muchos clientes optimizan su objetivo de tiempo de recuperación al ampliar un clúster de Oracle RAC en todos los sitios, lo que proporciona una configuración completamente activo-activo. El diseño general se complica porque debe incluir la gestión de quórum de Oracle RAC. Además, se accede a los datos desde ambos sitios, lo que significa que una conmutación por error forzada puede provocar el uso de una copia desactualizada de los datos.

Aunque se encuentra una copia de los datos en ambos sitios, solo la controladora que actualmente posee un agregado puede servir datos. Por lo tanto, con clústeres RAC ampliados, los nodos remotos deben ejecutar operaciones de I/O a través de una conexión de sitio a sitio. El resultado es una latencia de I/O añadida, pero esta latencia no suele ser un problema. La red de interconexión de RAC también debe extenderse entre sitios, lo que significa que se necesita una red de alta velocidad y baja latencia de todos modos. Si la latencia añadida provoca un problema, el clúster se puede operar de una forma activa-pasiva. Luego, las operaciones con un gran volumen de I/O deben dirigirse a los nodos de RAC locales a la controladora propietaria de los agregados. A continuación, los nodos remotos realizan operaciones de E/S más ligeras o se utilizan únicamente como servidores de espera templados.

Si se requiere un RAC extendido activo-activo, se debe considerar el mirroring de ASM en lugar de MetroCluster. La duplicación de ASM permite que se prefiera una réplica específica de los datos. Por lo tanto, se puede crear un clúster RAC ampliado en el que todas las lecturas se realicen localmente. La I/O de lectura nunca se cruza con los sitios, lo que ofrece la menor latencia posible. Toda la actividad de escritura debe seguir transfiriendo la conexión entre sitios, pero dicho tráfico es inevitable con cualquier solución de mirroring síncrono.

Si las LUN de inicio, incluidos los discos de inicio virtualizados, se utilizan con Oracle RAC, el `misscount` es posible que sea necesario cambiar el parámetro. Para obtener más información sobre los parámetros de tiempo de espera de RAC, consulte "Oracle RAC con ONTAP".

Configuración de dos sitios

Una configuración de RAC ampliada de dos sitios puede ofrecer servicios de base de datos activa-activa que pueden sobrevivir muchos escenarios de desastres de forma no disruptiva, pero no todos.

Archivos de quorum de RAC

La primera consideración al implementar RAC ampliado en MetroCluster debe ser la gestión del quórum. Oracle RAC tiene dos mecanismos para gestionar el quórum: Latido de disco y latido de red. El latido del disco supervisa el acceso al almacenamiento mediante los archivos de votación. Con una configuración de RAC de un único sitio, un único recurso de votación es suficiente siempre que el sistema de almacenamiento subyacente ofrezca funcionalidades de alta disponibilidad.

En versiones anteriores de Oracle, los archivos de quorum se colocaban en dispositivos de almacenamiento físico, pero en las versiones actuales de Oracle los archivos de quorum se almacenan en grupos de discos de ASM.

Oracle RAC es compatible con NFS. Durante el proceso de instalación de grid, se crea un juego de procesos de ASM para presentar la ubicación NFS utilizada para los archivos de grid como un grupo de discos de ASM. El proceso es prácticamente transparente para el usuario final y no requiere una gestión de ASM en curso una vez finalizada la instalación.

El primer requisito de una configuración de dos ubicaciones es asegurarse de que cada sitio siempre pueda acceder a más de la mitad de los archivos de votación de forma que se garantice un proceso de recuperación ante desastres sin interrupciones. Esta tarea era sencilla antes de que los archivos de votación se almacenaran en grupos de discos de ASM, pero hoy en día los administradores necesitan comprender los principios básicos de la redundancia de ASM.

Los grupos de discos de ASM tienen tres opciones de redundancia `external`, `normal`, y. `high`. En otras palabras, se refleja en 3 direcciones y no reflejado. Una opción más reciente llamada `Flex` también está disponible, pero rara vez se utiliza. El nivel de redundancia y la ubicación de los dispositivos redundantes controlan lo que sucede en escenarios de fallo. Por ejemplo:

- Colocación de los archivos de votación en un `diskgroup` con `external` los recursos de redundancia garantizan el desalojo de un sitio si se pierde la conectividad entre sitios.
- Colocación de los archivos de votación en un `diskgroup` con `normal` La redundancia con un solo disco ASM por sitio garantiza la expulsión de nodos en ambas ubicaciones si se pierde la conectividad entre sitios porque ninguno de los sitios tendría un quórum mayoritario.
- Colocación de los archivos de votación en un `diskgroup` con `high` la redundancia con dos discos en un sitio y un solo disco en el otro sitio permite las operaciones activo-activo cuando ambos sitios están operativos y se puede acceder mutuamente. Sin embargo, si el sitio de un solo disco está aislado de la red, ese sitio se expulsa.

Latido de red RAC

El latido de red de Oracle RAC supervisa la accesibilidad de nodos en la interconexión de cluster. Para permanecer en el clúster, un nodo debe ser capaz de contactar más de la mitad de los otros nodos. En una arquitectura de dos sitios, este requisito crea las siguientes opciones para el recuento de nodos de RAC:

- La colocación de un número igual de nodos por sitio provoca el expulsión en un sitio en caso de que se pierda la conectividad de red.
- La colocación de los nodos N en un sitio y los nodos N+1 en el sitio opuesto garantiza que la pérdida de

conectividad entre sitios da lugar al sitio con el mayor número de nodos restantes en el quórum de red y el sitio con menos nodos expulsados.

Antes de Oracle 12cR2, no era posible controlar qué lado experimentaría un desalojo durante la pérdida del sitio. Cuando cada ubicación tiene el mismo número de nodos, el nodo maestro controla la expulsión, que en general es el primer nodo RAC que se inicia.

Oracle 12cR2 introduce la capacidad de ponderación de nodos. Esta capacidad proporciona al administrador más control sobre cómo Oracle resuelve las condiciones de cerebro dividido. Como ejemplo sencillo, el siguiente comando establece la preferencia de un nodo concreto en 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.
```

Después de reiniciar Oracle High-Availability Services, la configuración tiene el siguiente aspecto:

```
[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` ahora se designa como servidor crítico. Si los dos nodos de RAC están aislados, `host-a` sobrevive, y. `host-b` se expulsa.

Para obtener más información, consulte el white paper de Oracle sobre Oracle Clusterware 12c Versión 2 Technical Overview. ”

Para las versiones de Oracle RAC anteriores a 12cR2, el nodo maestro se puede identificar comprobando los logs de CRS de la siguiente manera:

```
[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
```

Este log indica que el nodo maestro es `2` y el nodo `host-a` Tiene un ID de `1`. Este hecho significa eso `host-a` no es el nodo maestro. La identidad del nodo maestro se puede confirmar con el comando `olsnodes -n`.

```
[root@host-a ~]# /grid/bin/olsnodes -n
host-a  1
host-b  2
```

El nodo con un ID de `2` es `host-b`, que es el nodo maestro. En una configuración con el mismo número de nodos en cada sitio, el sitio con `host-b` es el sitio que sobrevive si los dos conjuntos pierden la conectividad de red por cualquier motivo.

Es posible que la entrada de log que identifica el nodo maestro pueda quedar obsoleta en el sistema. En esta situación, se pueden utilizar las marcas de tiempo de las copias de seguridad de Oracle Cluster Registry (OCR).

```
[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
```

En este ejemplo se muestra que el nodo maestro es `host-b`. También indica un cambio en el nodo maestro desde `host-a` para `host-b` En algún lugar entre las 2:05 y las 21:39 el 4 de mayo. Este método de

identificación del nodo maestro sólo es seguro si también se han comprobado los registros de CRS porque es posible que el nodo maestro haya cambiado desde la copia de seguridad de OCR anterior. Si se ha producido este cambio, debería estar visible en los registros de OCR.

La mayoría de los clientes eligen un único grupo de discos de votación que da servicio a todo el entorno y un número igual de nodos de RAC en cada sitio. El grupo de discos se debe colocar en el sitio que contiene la base de datos. El resultado es que la pérdida de conectividad provoca el desalojo en el sitio remoto. El sitio remoto ya no tendría quórum ni tendría acceso a los archivos de la base de datos, pero el sitio local continúa funcionando como de costumbre. Cuando se restaura la conectividad, la instancia remota puede volver a conectarse.

En caso de desastre, se requiere un cambio para poner los archivos de la base de datos y el grupo de discos de votación en línea en el sitio superviviente. Si el desastre permite que AUSO active la conmutación por error, NVFAIL no se activa porque se sabe que el clúster está sincronizado y que los recursos de almacenamiento se conectan de forma normal. AUSO es una operación muy rápida y debe completarse antes de la `disktimeout` el período caduca.

Dado que solo hay dos sitios, no es factible utilizar ningún tipo de software automatizado de tiebreaking externo, lo que significa que la conmutación por error forzada debe ser una operación manual.

Configuraciones en tres sitios

Un clúster RAC ampliado es mucho más fácil de diseñar con tres sitios. Los dos sitios que alojan cada mitad del sistema de MetroCluster también admiten cargas de trabajo de base de datos, mientras que el tercer sitio sirve como desempate tanto para la base de datos como para el sistema de MetroCluster. La configuración de Oracle tiebreaker puede ser tan sencilla como colocar un miembro del grupo de discos de ASM utilizado para votar en un sitio 3rd y también puede incluir una instancia operativa en el sitio 3rd para asegurarse de que hay un número impar de nodos en el cluster RAC.

Consulte la documentación de Oracle sobre el “grupo de fallos de quórum” para obtener información importante sobre el uso de NFS en una configuración RAC ampliada. En resumen, puede que sea necesario modificar las opciones de montaje NFS para incluir la opción soft para garantizar que la pérdida de conectividad con los recursos de quórum del sitio de 3rd que alojan no cuelgue los servidores Oracle principales ni los procesos de Oracle RAC.

Información de copyright

Información de la marca comercial