Recuperación ante desastres de Oracle

Enterprise applications

NetApp
February 10, 2026

Tabla de contenidos

Recuperación ante desastres de Oracle

Descripción general

La recuperación tras desastres se refiere a la restauración de servicios de datos tras un evento catastrófico, como un incendio que destruye un sistema de almacenamiento o incluso un sitio entero.

Esta documentación sustituye a los informes técnicos publicados anteriormente *TR-4591: Oracle Data Protection* y *TR-4592: Oracle en MetroCluster.*

La recuperación tras desastres puede llevarse a cabo mediante la replicación sencilla de datos mediante SnapMirror; por supuesto, muchos clientes actualizan réplicas replicadas cada hora.

Para la mayoría de los clientes, la recuperación ante desastres requiere algo más que poseer una copia remota de datos, requiere la capacidad para usar rápidamente esos datos. NetApp ofrece dos tecnologías para satisfacer esta necesidad: La sincronización activa de MetroCluster y SnapMirror

MetroCluster se refiere a ONTAP en una configuración de hardware que incluye almacenamiento reflejado sincrónico de bajo nivel y numerosas funciones adicionales. Las soluciones integradas como MetroCluster simplifican las complicadas infraestructuras de bases de datos, aplicaciones y virtualización actuales y de escalado horizontal. Reemplaza múltiples productos y estrategias de protección de datos externa por una cabina de almacenamiento simple y central. También proporciona integración de backup, recuperación, recuperación tras siniestros y alta disponibilidad (HA) en un único sistema de almacenamiento en clúster.

La sincronización activa (SM-AS) de SnapMirror se basa en SnapMirror síncrono. Con MetroCluster, cada controladora de ONTAP es responsable de replicar los datos de la unidad en una ubicación remota. Con la sincronización activa de SnapMirror, básicamente cuenta con dos sistemas ONTAP diferentes que mantienen copias independientes de los datos de su unidad lógica, pero que cooperan para presentar una única instancia de esa LUN. Desde el punto de vista del host, se trata de una única entidad de LUN.

Comparación SM-AS y MCC

SM-AS y MetroCluster son similares en cuanto a funcionalidad general, pero hay diferencias importantes en la forma en que se implementó la replicación RPO=0 y cómo se gestiona. La sincronización asíncrona y síncrona de SnapMirror también se puede utilizar como parte de un plan de recuperación ante desastres, pero no están diseñadas como tecnologías de réplica de alta disponibilidad.

- Una configuración de MetroCluster es más como un clúster integrado con nodos distribuidos por todos los sitios. SM-AS se comporta como dos clústeres independientes que cooperan en el servicio de un objetivo de punto de recuperación seleccionado=0 LUN replicadas de forma síncrona.
- Los datos de una configuración MetroCluster solo son accesibles desde un sitio concreto en un momento dado. Una segunda copia de los datos está presente en el sitio opuesto, pero los datos son pasivos. No es posible acceder a ella sin una conmutación por error del sistema de almacenamiento.
- El mirroring de MetroCluster y SM-AS se produce en diferentes niveles. El mirroring de MetroCluster se realiza en la capa de RAID. Los datos de bajo nivel se almacenan en un formato duplicado mediante SyncMirror. El uso del mirroring es prácticamente invisible en las capas de LUN, volumen y protocolo.
- Por el contrario, la duplicación SM-AS se produce en la capa de protocolo. Los dos clústeres son clústeres independientes en general. Una vez que las dos copias de datos están sincronizadas, los dos clústeres solo tienen que reflejar las escrituras. Cuando se produce una escritura en un clúster, se replica en otro

clúster. La escritura solo se reconoce en el host cuando la escritura se ha completado en ambos sitios. Aparte de este comportamiento de división del protocolo, los dos clústeres son clústeres ONTAP normales.

- El rol principal de MetroCluster es la replicación a gran escala. Puede replicar toda una cabina con RPO=0 y RTO casi nulo. Esto simplifica el proceso de conmutación al nodo de respaldo porque solo hay una cosa que conmutar al nodo de respaldo y ofrece una escalabilidad extremadamente buena en cuanto a capacidad e IOPS.
- Un caso de uso clave de SM-AS es la replicación granular. En ocasiones, no desea replicar todos los datos como una sola unidad, o debe poder conmutar al nodo de respaldo de forma selectiva ciertas cargas de trabajo.
- Otro caso de uso clave de SM-AS es en las operaciones activo-activo, donde desea que existan copias de datos totalmente utilizables para estar disponibles en dos clústeres diferentes ubicados en dos ubicaciones diferentes con idénticas características de rendimiento y, si lo desea, no es necesario ampliar la SAN entre los sitios. Puede tener sus aplicaciones ya ejecutándose en ambos sitios, lo que reduce el RTO general durante las operaciones de conmutación al respaldo.

MetroCluster

Recuperación ante desastres con MetroCluster

MetroCluster es una función de ONTAP que puede proteger sus bases de datos de Oracle con RPO=0 mirroring sincrónico entre sitios y se escala verticalmente para admitir cientos de bases de datos en un único sistema de MetroCluster.

También es fácil de usar. El uso de MetroCluster no es necesariamente uno de los factores que contribuyen a aumentar ni cambiar los mejores cursos para el funcionamiento de aplicaciones y bases de datos empresariales.

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.

Arquitectura física

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

El término 'conectividad' hace referencia a la conexión de clúster usada 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.

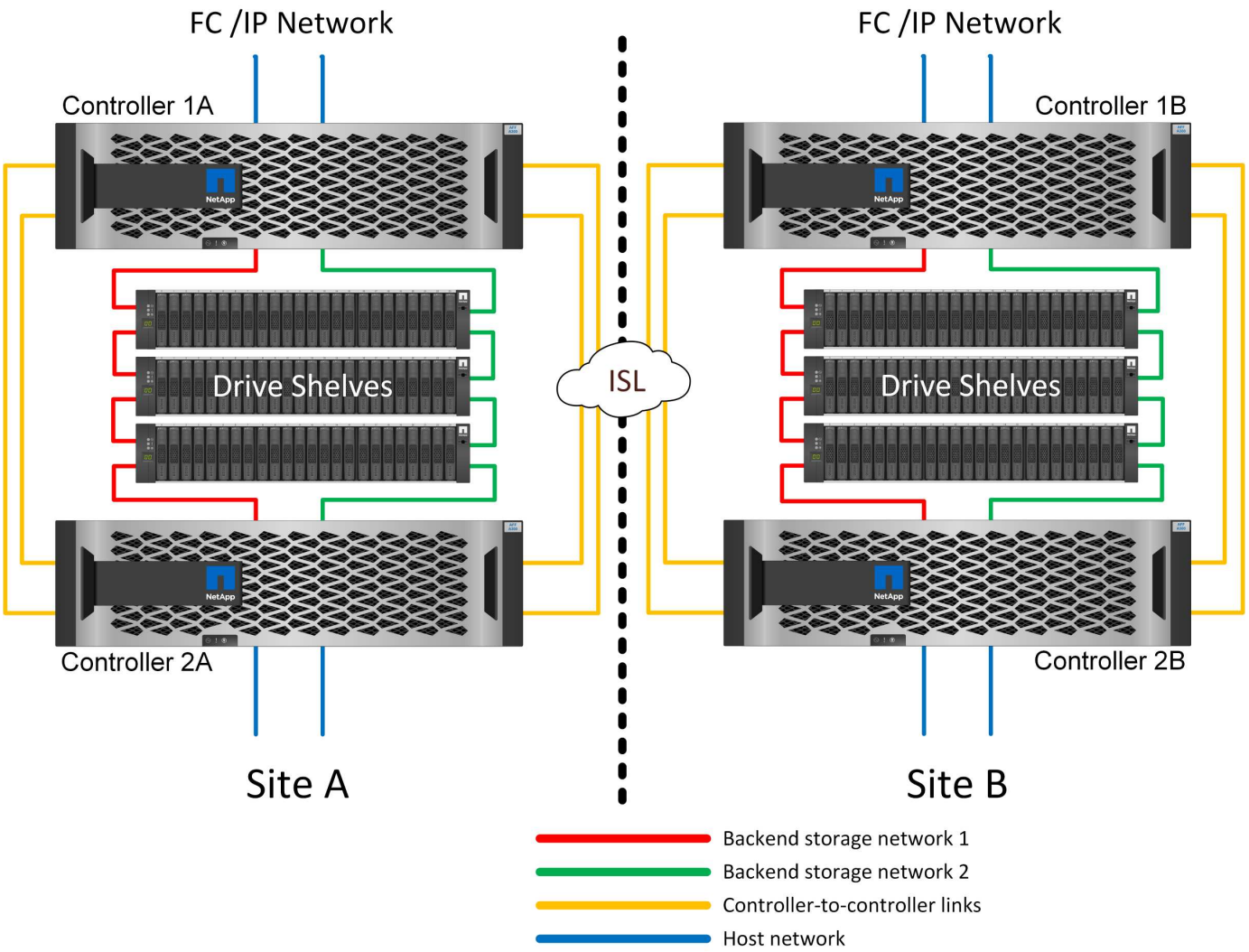
IP de MetroCluster

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.

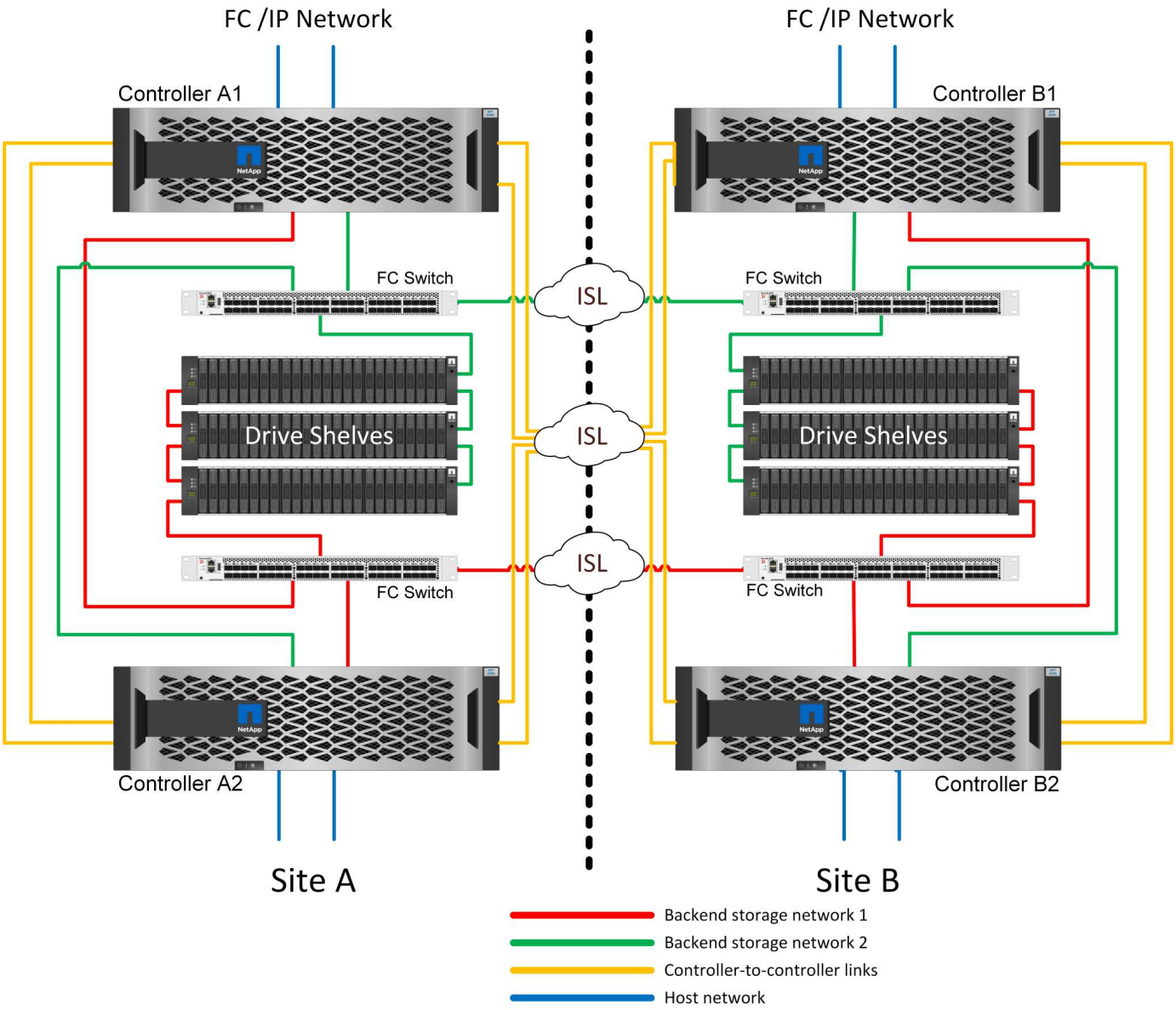
La arquitectura además es más sencilla ya que las únicas conexiones entre sitios son para las controladoras. En MetroCluster conectados a FC SAN, una controladora escribe directamente en las unidades del sitio opuesto y, por lo tanto, requiere conexiones SAN, switches y puentes adicionales. En cambio,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



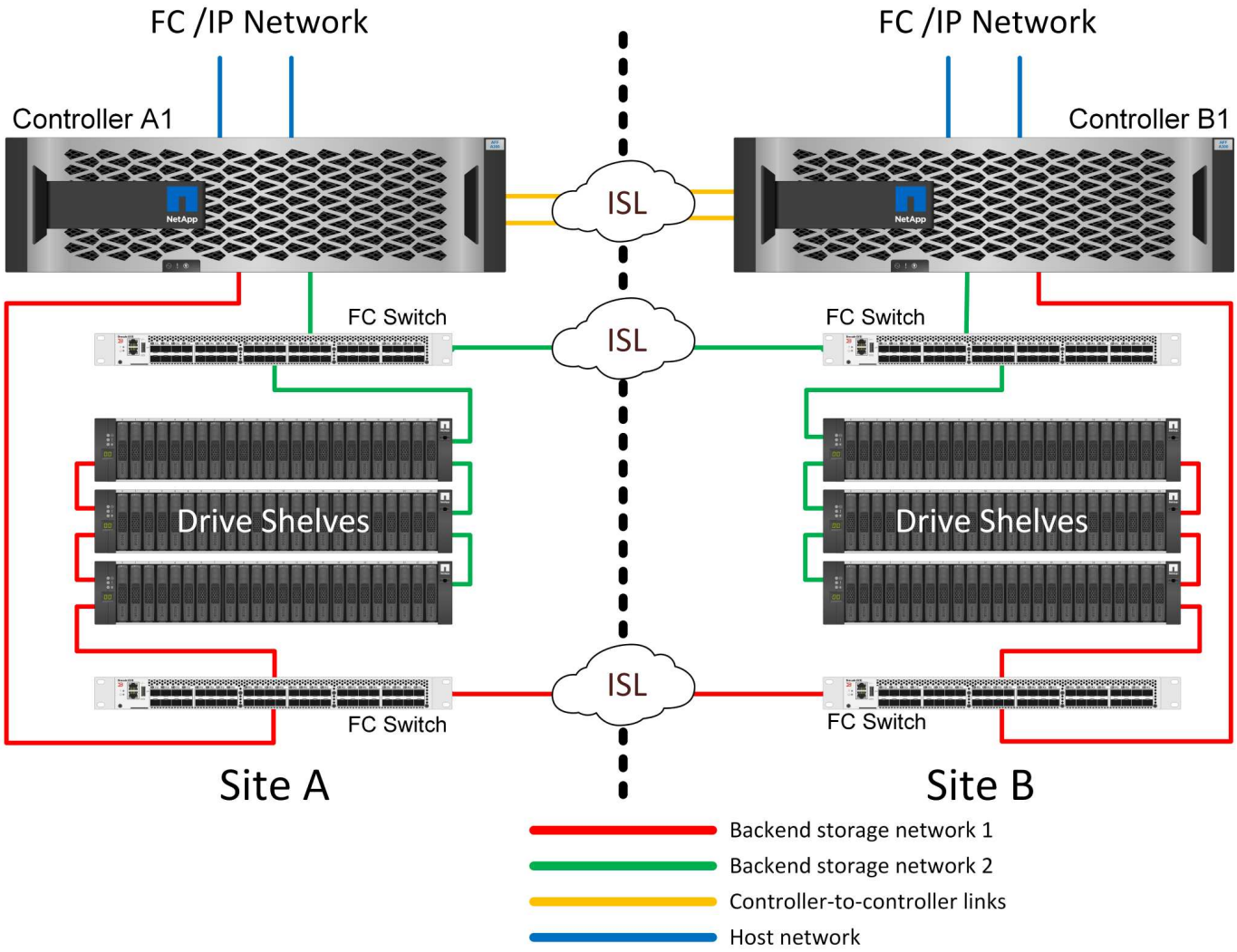
Algunas infraestructuras multisitio no están diseñadas para operaciones activo-activo, sino que se utilizan más como sitio principal y sitio de recuperación de desastres. En esta situación, generalmente es preferible una opción MetroCluster de una pareja de alta disponibilidad por las siguientes razones:

- 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

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

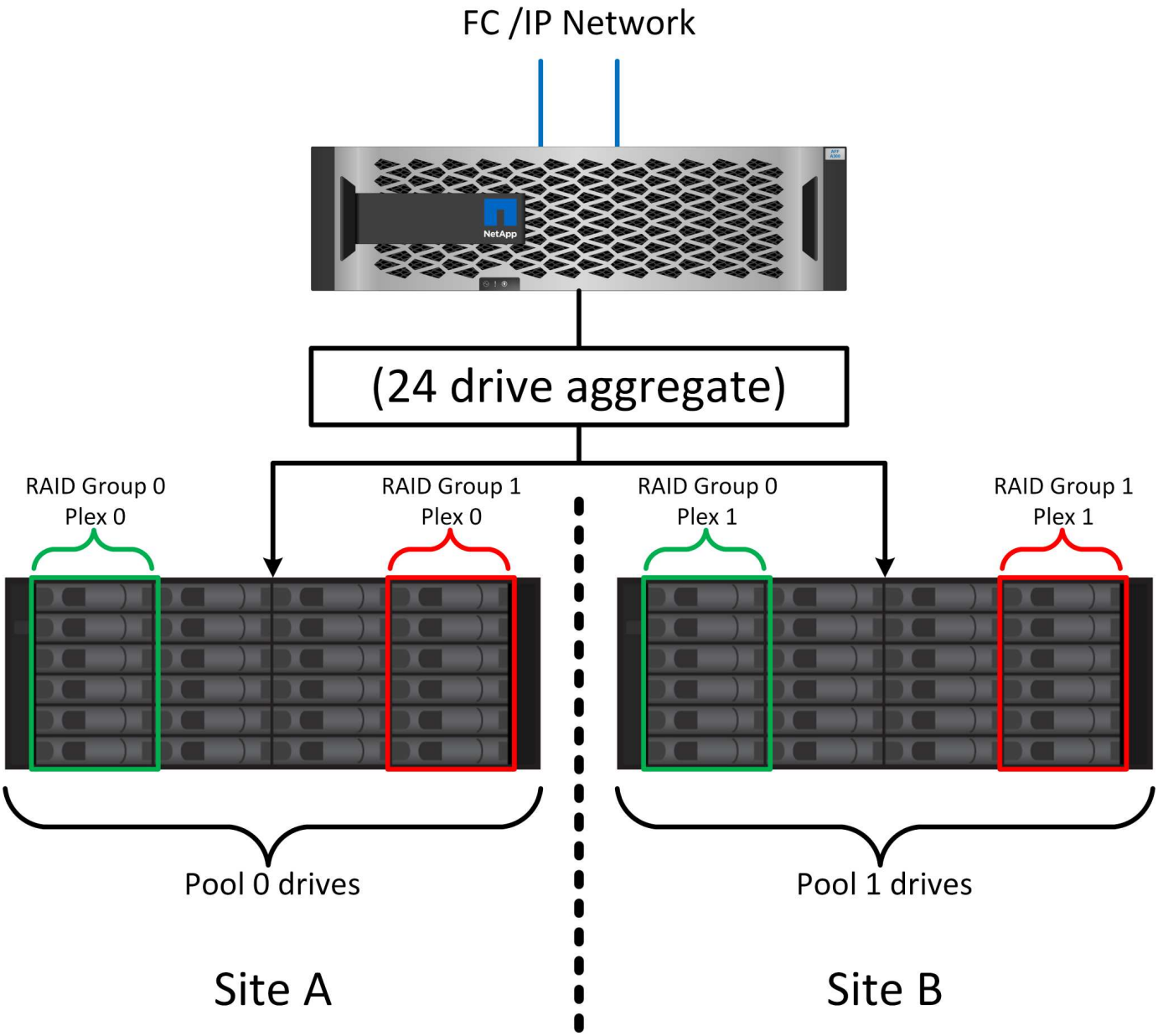
MetroCluster amplía la protección de datos de NVRAM de las siguientes formas:

- 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

"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 principal objetivo de MetroCluster tiebreaker es detectar la pérdida de sitios. 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.

> El software MCTB no verifica que NVRAM WAS y/o los plexes estén sincronizados al forzar un switchover. 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, al igual que el software MetroCluster tiebreaker de referencia anteriormente, pero también incluye una característica crítica, con la posibilidad de realizar una conmutación de sitios automatizada sin supervisión.

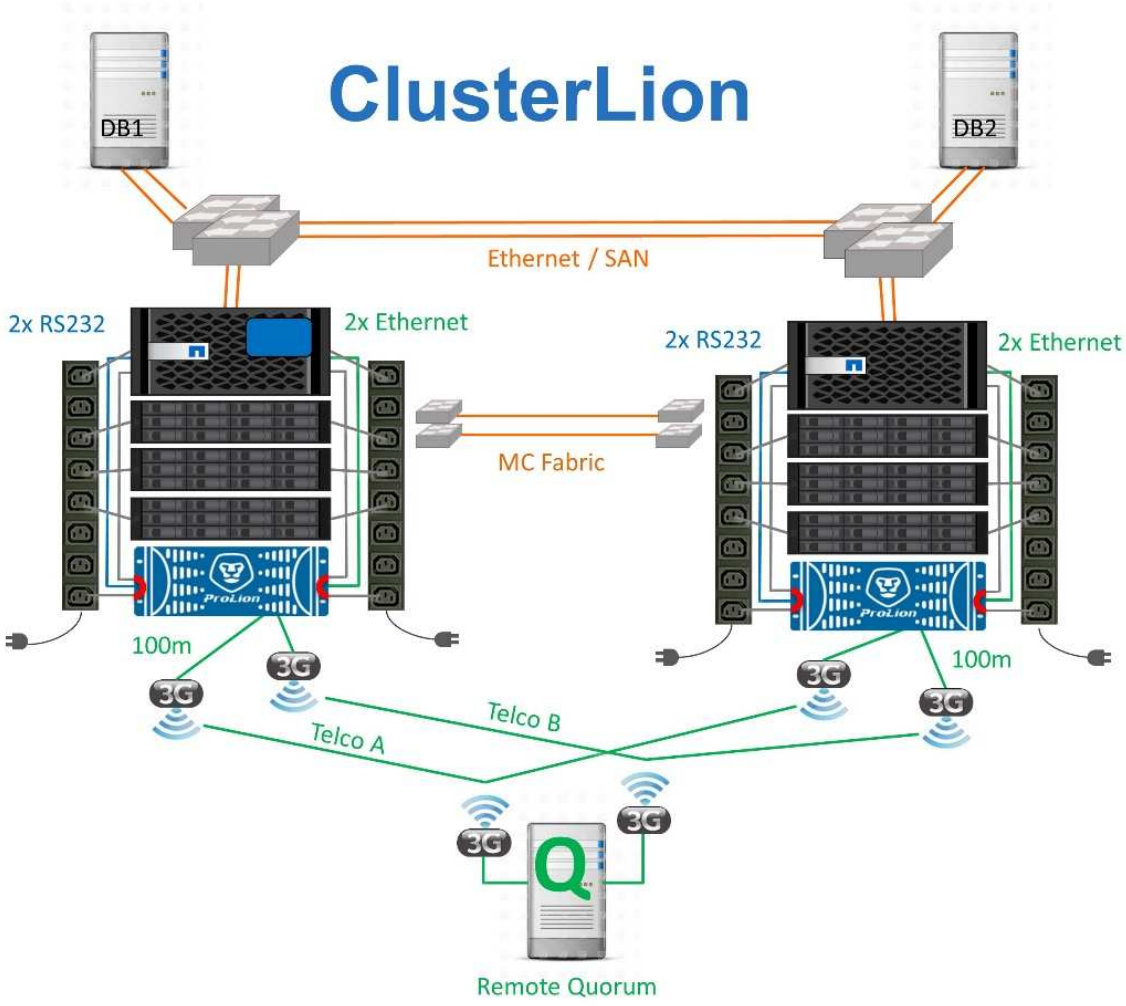
Una MetroCluster conectada a estructura tiene acceso directo a dispositivos de almacenamiento en el sitio opuesto. Esto permite que una controladora MetroCluster supervise el estado de las otras controladoras mediante la lectura de datos de latidos de las unidades. Esto permite que una controladora reconozca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

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 SQL Server, grupos de disponibilidad Always On. El mirroring de discos a nivel de sistema operativo puede ser una opción. Consulte con su equipo de cuentas de partner o de NetApp para obtener más información y opciones.

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

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 Extended RAC

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 la sincronización activa de SnapMirror en lugar de MetroCluster. La replicación de SM-AS 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 se utilizan LUN de inicio, incluidos los discos de inicio virtualizados, con Oracle RAC, es posible que el `misscount` parámetro deba cambiarse. Para obtener más información sobre los parámetros de timeout 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.

SnapMirror síncrono activo

Descripción general

SnapMirror Active Sync le permite crear entornos de base de datos de Oracle de alta disponibilidad donde los LUN están disponibles desde dos clústeres de almacenamiento diferentes.

Con la sincronización activa de SnapMirror, no hay copias «primarias» ni «secundarias» de los datos. Cada clúster puede servir de I/O de lectura a partir de su copia local de los datos y cada clúster replicará una escritura en su compañero. El resultado es un comportamiento de E/S simétrico.

Entre otras opciones, esto permite ejecutar Oracle RAC como un cluster ampliado con instancias operativas en ambas ubicaciones. También podría crear un RPO=0 clusters de bases de datos activo-pasivo en los que las bases de datos de una instancia única se puedan mover de un sitio a otro durante una interrupción del servicio. Este proceso puede automatizarse mediante productos como Pacemaker o VMware HA. La base de todas estas opciones es la replicación síncrona gestionada por SnapMirror Active Sync.

Replicación síncrona

En un funcionamiento normal, la sincronización activa de SnapMirror proporciona una réplica síncrona con un objetivo de punto de recuperación=0 en todo momento, con una excepción. Si los datos no se pueden replicar, ONTAP liberará el requisito para replicar datos y reanudar el servicio de I/O en un sitio mientras las LUN del otro sitio se desconecten.

Hardware de almacenamiento

Al contrario que otras soluciones de recuperación ante desastres del almacenamiento, SnapMirror Active Sync ofrece una flexibilidad de plataforma asimétrica. No es necesario que el hardware de cada sitio sea idéntico. Esta funcionalidad permite ajustar el tamaño adecuado del hardware que se utiliza para dar soporte a SnapMirror de sincronización activa. El sistema de almacenamiento remoto puede ser idéntico al sitio principal si necesita soportar una carga de trabajo de producción completa, pero si un desastre provoca una reducción de I/O, es posible que un sistema más pequeño en el sitio remoto sea más rentable.

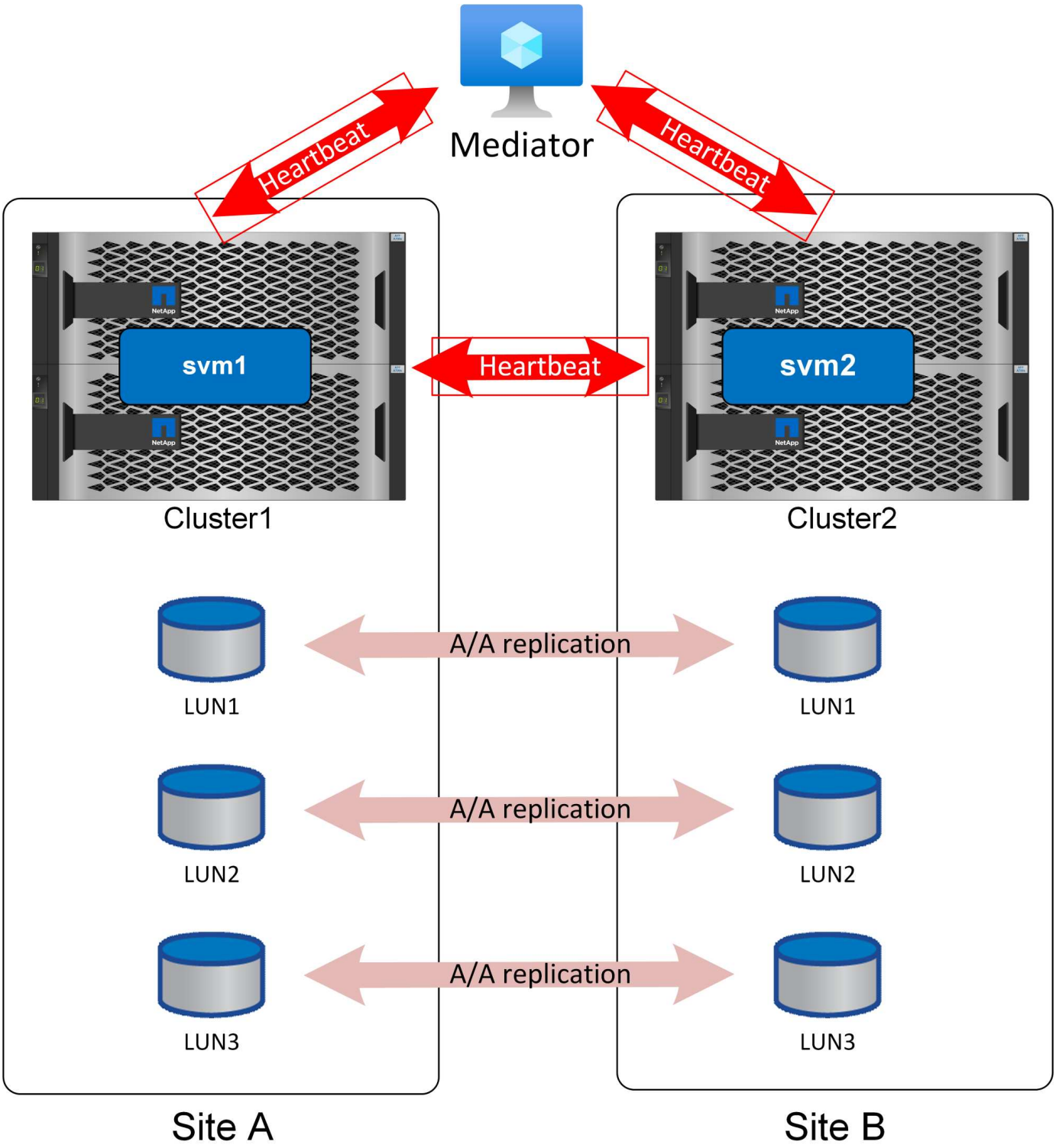
Mediador ONTAP

ONTAP Mediator es una aplicación de software que se descarga del soporte técnico de NetApp y que normalmente se implementa en una pequeña máquina virtual. ONTAP Mediator no es un tiebreaker cuando se utiliza con SnapMirror sincronización activa. Es un canal de comunicación alternativo para los dos clústeres que participan en la replicación síncrona activa de SnapMirror. Las operaciones automatizadas son impulsadas por ONTAP en función de las respuestas recibidas del partner a través de conexiones directas y a través del mediador.

Mediador ONTAP

El mediador es necesario para automatizar la conmutación por error de forma segura. Lo ideal sería que se ubicara en un sitio 3rd independiente, pero todavía puede funcionar para la mayoría de las necesidades si se ubicara con uno de los clústeres que participan en la replicación.

El mediador no es realmente un desempate, aunque esa sea, en efecto, la función que cumple. El mediador ayuda a determinar el estado de los nodos del clúster y asiste en el proceso de conmutación automática en caso de fallo del sitio. El mediador no transfiere datos bajo ninguna circunstancia.



El desafío #1 con la conmutación automática por error es el problema de cerebro dividido, y ese problema surge si sus dos sitios pierden conectividad entre sí. ¿Qué debería pasar? No desea que dos sitios diferentes se designen a sí mismos como copias supervivientes de los datos, pero ¿cómo puede un solo sitio diferenciar entre la pérdida real del sitio opuesto y la incapacidad de comunicarse con el sitio opuesto?

Aquí es donde el mediador entra en la imagen. Si se coloca en un sitio 3rd y cada sitio tiene una conexión de red independiente con ese sitio, tiene una ruta adicional para que cada sitio valide el estado del otro. Mire la imagen de arriba otra vez y considere los siguientes escenarios.

- ¿Qué sucede si el mediador falla o es inaccesible desde uno o ambos sitios?
  - Los dos clústeres pueden comunicarse entre sí a través del mismo enlace utilizado para los servicios de replicación.
  - Los datos se siguen ofreciendo con la protección RPO=0
- ¿Qué sucede si falla el sitio A?
  - El sitio B verá que ambos canales de comunicación se caen.
  - El sitio B se hará cargo de los servicios de datos, pero sin el mirroring RPO=0
- ¿Qué sucede si falla el sitio B?
  - El sitio A verá que ambos canales de comunicación se caen.
  - El sitio A se hará cargo de los servicios de datos, pero sin el mirroring RPO=0

Hay otro escenario a considerar: Pérdida del enlace de replicación de datos. Si se pierde el enlace de replicación entre los sitios, obviamente el mirroring RPO=0 se convertirá en imposible. ¿Qué debería pasar entonces?

Esto se controla por el estado de sitio preferido. En una relación SM-AS, uno de los sitios es secundario al otro. Esto no afecta a las operaciones normales y todo el acceso a los datos es simétrico, pero si la replicación se interrumpe, el vínculo tendrá que romperse para reanudar las operaciones. Como resultado, el sitio preferido continuará con las operaciones sin mirroring y el sitio secundario detendrá el procesamiento de I/O hasta que se restaure la comunicación de la replicación.

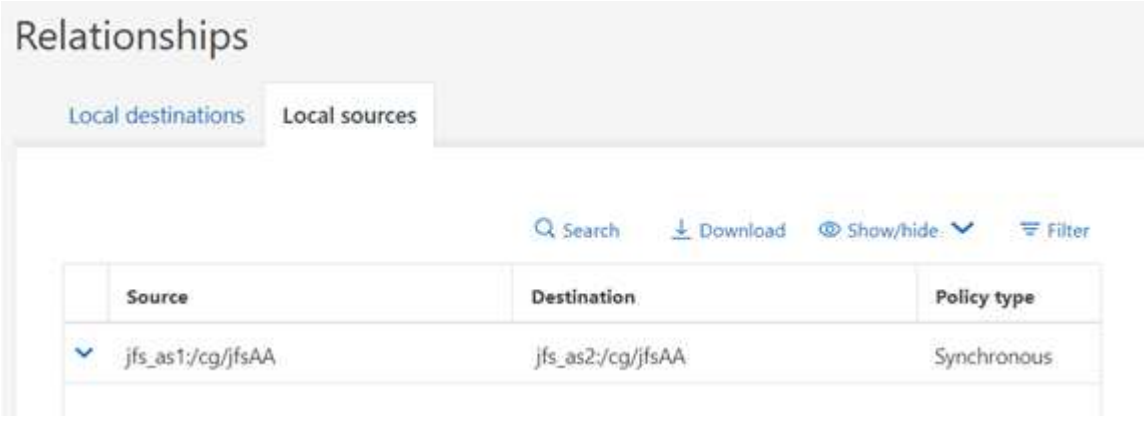
Sitio preferido de sincronización activa de SnapMirror

El comportamiento de sincronización activa de SnapMirror es simétrico, con una excepción importante: La configuración de sitio preferida.

La sincronización activa de SnapMirror considerará que un sitio es el «origen» y el otro el «destino». Esto implica una relación de replicación unidireccional, pero esto no se aplica al comportamiento de E/S. La replicación es bidireccional y simétrica, y los tiempos de respuesta de I/O son los mismos en cualquier lado del espejo.

La `source` designación controla el sitio preferido. Si se pierde el enlace de replicación, las rutas de LUN en la copia de origen seguirán sirviendo datos mientras las rutas de LUN en la copia de destino dejarán de estar disponibles hasta que se restablezca la replicación y SnapMirror vuelva a entrar en un estado síncrono. A continuación, las rutas reanudarán el servicio de datos.

La configuración de origen/destino se puede ver a través de SystemManager:



O en la CLI:

```
Cluster2::> snapmirror show -destination-path jfs_as2:/cg/jfsAA

                           Source Path: jfs_as1:/cg/jfsAA
                     Destination Path: jfs_as2:/cg/jfsAA
                    Relationship Type: XDP
                Relationship Group Type: consistencygroup
                    SnapMirror Schedule: -
                 SnapMirror Policy Type: automated-failover-duplex
                      SnapMirror Policy: AutomatedFailOverDuplex
                            Tries Limit: -
                      Throttle (KB/sec): -
                           Mirror State: Snapmirrored
                    Relationship Status: InSync
```

La clave es que la fuente es la máquina virtual de almacenamiento SVM en cluster1. Tal como se ha mencionado anteriormente, los términos «origen» y «destino» no describen el flujo de los datos replicados. Ambos sitios pueden procesar una escritura y replicarla en el sitio opuesto. De hecho, ambos clústeres son orígenes y destinos. El efecto de designar un clúster como origen simplemente controla qué clúster sobrevive como sistema de almacenamiento de lectura y escritura si se pierde el enlace de replicación.

Topología de red

Acceso uniforme

La conexión de red de acceso uniforme significa que los hosts pueden acceder a las rutas en ambos sitios (o dominios de fallo dentro del mismo sitio).

Una característica importante de SM-AS es la capacidad de configurar los sistemas de almacenamiento para conocer dónde se encuentran los hosts. Cuando asigna las LUN a un host determinado, puede indicar si son proximales o no a un sistema de almacenamiento determinado.

Ajustes de proximidad

La proximidad se refiere a una configuración por clúster que indica que un ID de iniciador de iSCSI o WWN de host particular pertenece a un host local. Es un segundo paso opcional para configurar el acceso a LUN.

El primer paso es la configuración habitual del igroup. Cada LUN debe asignarse a un igroup que contiene los ID WWN/iSCSI de los hosts que necesitan acceder a ese LUN. Este controla el host que tiene *access* a una LUN.

El segundo paso opcional es configurar la proximidad del host. Esto no controla el acceso, controla *priority*.

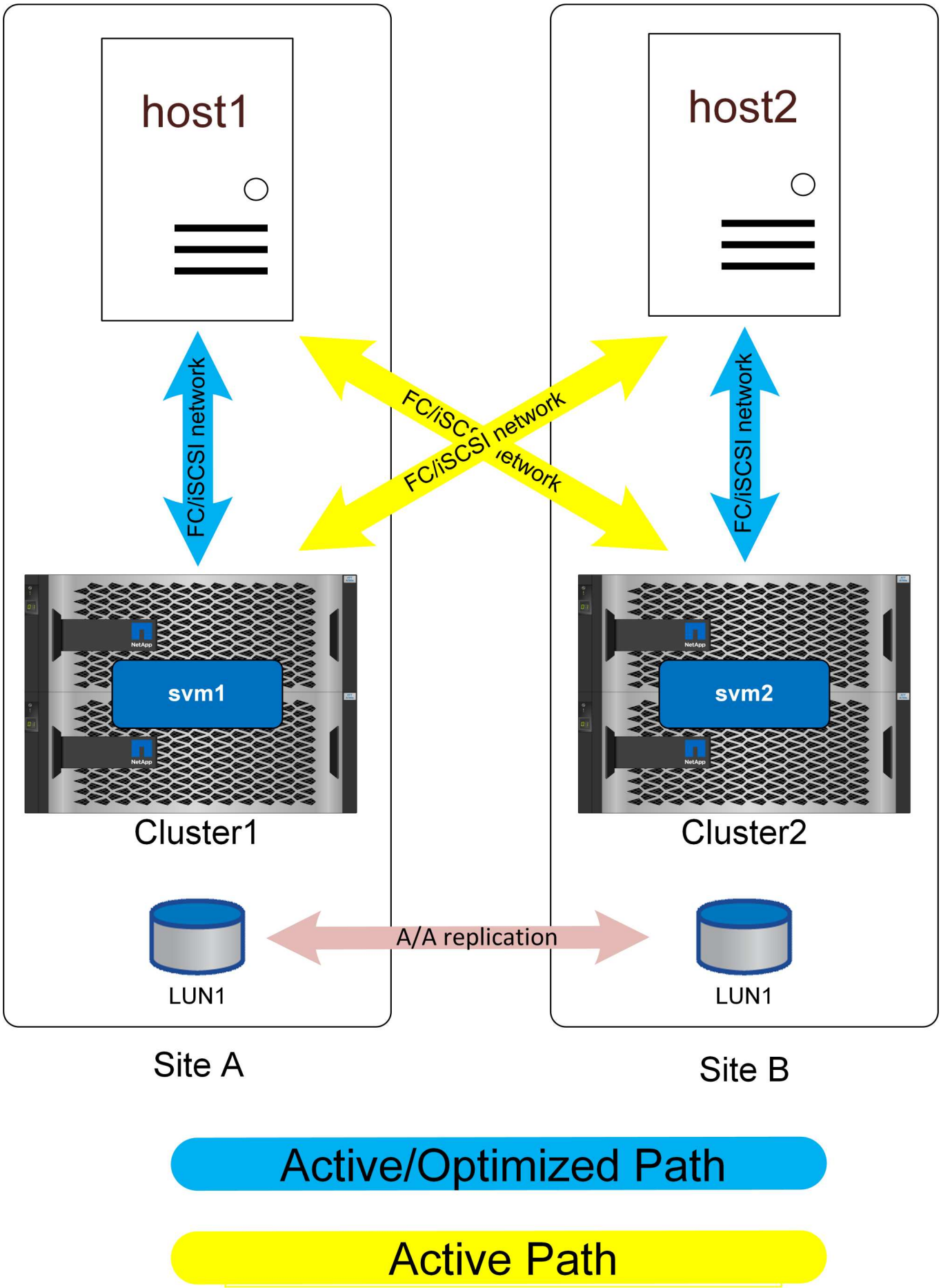
Por ejemplo, se puede configurar un host del sitio A para acceder a una LUN protegida por sincronización activa de SnapMirror y, como la SAN se extiende a través de los sitios, las rutas están disponibles para ese LUN usando el almacenamiento en el sitio A o el almacenamiento del sitio B.

Sin configuración de proximidad, ese host usará ambos sistemas de almacenamiento por igual porque ambos sistemas de almacenamiento anunciarán rutas activas/optimizadas. Si la latencia SAN o el ancho de banda

entre los sitios es limitada, es posible que no sea deseable y puede que desee asegurarse de que durante el funcionamiento normal cada host utilice preferentemente rutas hacia el sistema de almacenamiento local. Esto se configura añadiendo el ID de WWN/iSCSI de host al clúster local como un host proximal. Esto se puede hacer en la CLI o en SystemManager.

AFF

Con un sistema AFF, las rutas aparecerían como se muestra a continuación cuando se configura la proximidad del host.


host1
host2
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
svm1
svm2
Cluster1
Cluster2
A/A replication
LUN1
LUN1
Site A
Site B
Active/Optimized Path
Active Path

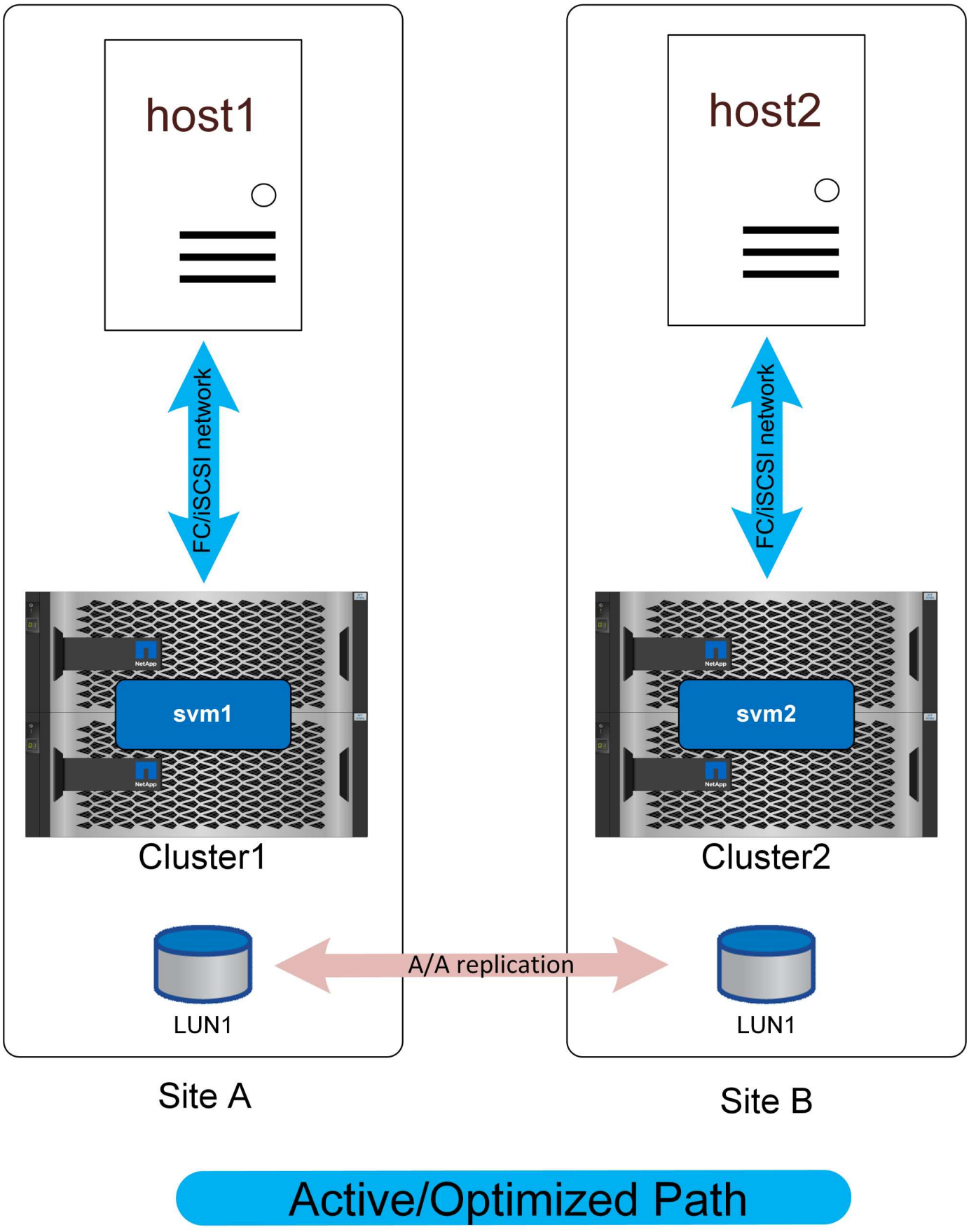
En funcionamiento normal, todas las I/O son locales. Las lecturas y escrituras se sirven desde la cabina de almacenamiento local. Por supuesto, el controlador local también deberá replicar el I/O de escritura en el sistema remoto antes de reconocerlo, pero todas las I/O de lectura se prestarán de servicio local y no incurrirán en latencia adicional atravesando el enlace SAN entre los sitios.

La única vez que se utilizarán las rutas no optimizadas es cuando se pierden todas las rutas activas/optimizadas. Por ejemplo, si se perdiera alimentación toda la cabina en el sitio A, los hosts del sitio A seguirían teniendo acceso a las rutas a la cabina en el sitio B y, por lo tanto, permanecerían operativos, aunque experimentarían una mayor latencia.

Existen rutas redundantes a través del clúster local que no se muestran en estos diagramas para simplificar el proceso. Los sistemas de almacenamiento de ONTAP son por sí mismos de alta disponibilidad, por lo que un fallo de una controladora no debe dar lugar a un fallo del sitio. Simplemente debe dar lugar a un cambio en el que se utilizan las rutas locales en el sitio afectado.

ASA

Los sistemas NetApp ASA ofrecen accesos múltiples activo-activo en todas las rutas de un clúster. Esto también se aplica a las configuraciones SM-AS.



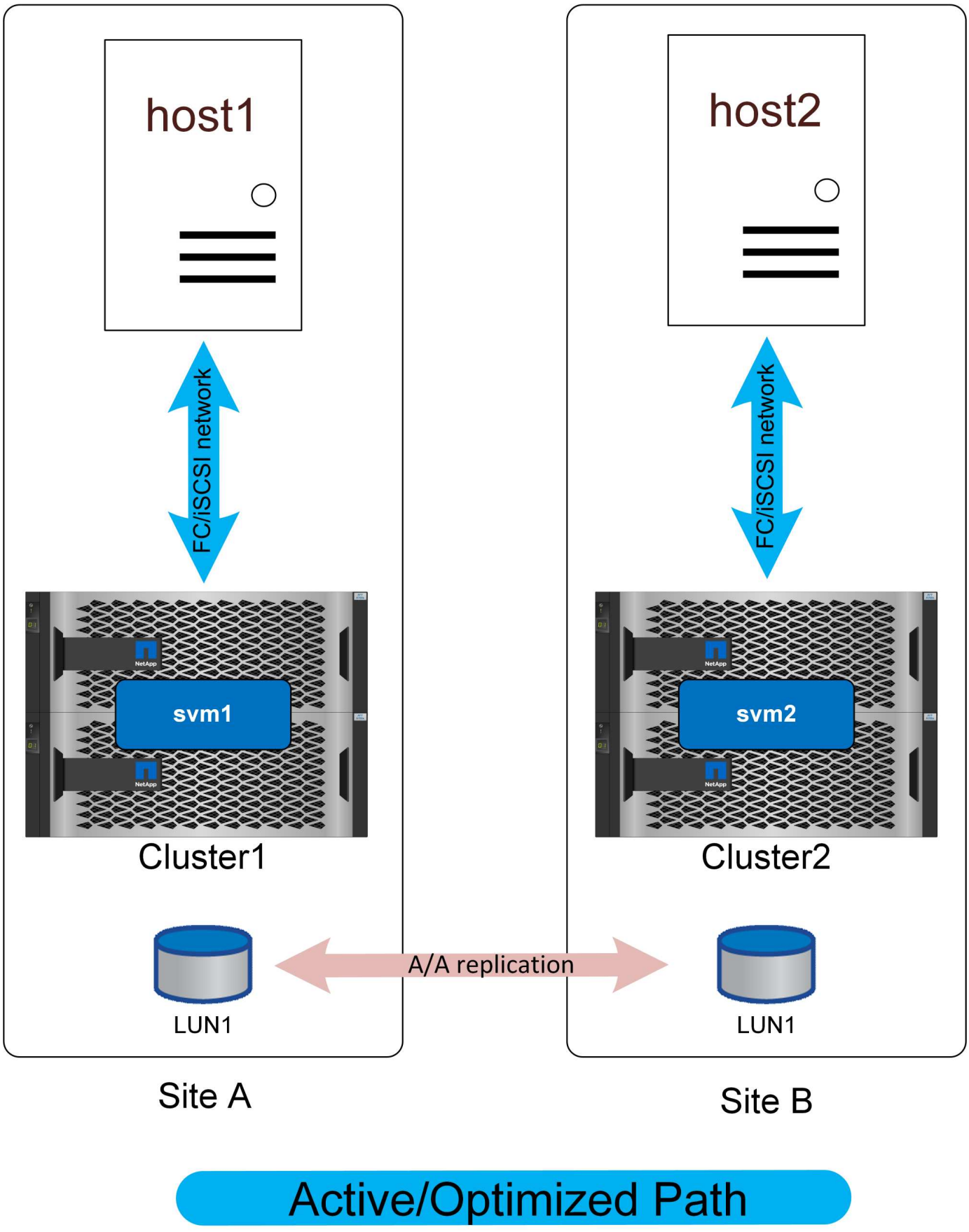
Una configuración de ASA con acceso no uniforme funcionaría en gran medida igual que con AFF. Con un acceso uniforme, IO estaría cruzando la WAN. Esto puede o no ser deseable.

Si los dos sitios estuvieran separados por 100 metros con conectividad de fibra, no debería haber una latencia adicional detectable que cruce la WAN, pero si los sitios estuvieran separados por una distancia larga, el rendimiento de lectura se vería afectado en ambos sitios. Por el contrario, con AFF, dichas rutas de cruce de WAN solo se utilizarían si no hubiera rutas locales disponibles y mejoraría el rendimiento diario, ya que todas las I/O serían locales. ASA con red de acceso no uniforme sería una opción para obtener las ventajas en coste y funciones de ASA sin incurrir en una penalización del acceso a la latencia de varios sitios.

ASA con SM en una configuración de baja latencia ofrece dos ventajas interesantes. En primer lugar, esencialmente * duplica * el rendimiento para cualquier host individual porque IO puede ser atendido por el doble de controladores usando el doble de rutas. En segundo lugar, en un entorno de sitio único ofrece una disponibilidad extrema debido a que se puede perder un sistema de almacenamiento completo sin interrumpir el acceso al host.

Acceso no uniforme

La red de acceso no uniforme significa que cada host solo tiene acceso a los puertos del sistema de almacenamiento local. La SAN no se extiende entre sitios (o dominios de fallos dentro del mismo sitio).



La principal ventaja de este método es la simplicidad en entornos SAN, eliminando la necesidad de extender una SAN por la red. Algunos clientes no tienen una conectividad de baja latencia suficiente entre los sitios o

no disponen de la infraestructura para túnel el tráfico de SAN FC a través de una red intersitio.

La desventaja del acceso no uniforme es que ciertos escenarios de fallo, incluida la pérdida del enlace de replicación, provocarán que algunos hosts pierdan el acceso al almacenamiento. Las aplicaciones que se ejecutan como instancias únicas, como una base de datos sin cluster que inherentemente solo se ejecuta en un único host en cualquier montaje dado, fallarían si se perdiera la conectividad del almacenamiento local. Los datos seguirían estando protegidos, pero el servidor de la base de datos ya no tendría acceso. Deberá reiniciarse en un sitio remoto, preferiblemente mediante un proceso automatizado. Por ejemplo, VMware HA puede detectar una situación de todas las rutas inactivas en un servidor y reiniciar una máquina virtual en otro servidor donde haya rutas disponibles.

Por el contrario, una aplicación en cluster como Oracle RAC puede ofrecer un servicio que esté disponible al mismo tiempo en dos sitios diferentes. Perder un sitio no significa la pérdida del servicio de la aplicación en su conjunto. Las instancias siguen estando disponibles y en ejecución en el sitio superviviente.

En muchos casos, sería inaceptable que la sobrecarga de latencia adicional derivada de una aplicación que accede al almacenamiento a través de un enlace entre sitio y sitio. Esto significa que la disponibilidad mejorada de una red uniforme es mínima, ya que la pérdida de almacenamiento en un sitio llevaría a la necesidad de apagar los servicios en ese sitio que ha fallado de todos modos.

Existen rutas redundantes a través del clúster local que no se muestran en estos diagramas para simplificar el proceso. Los sistemas de almacenamiento de ONTAP son por sí mismos de alta disponibilidad, por lo que un fallo de una controladora no debe dar lugar a un fallo del sitio. Simplemente debe dar lugar a un cambio en el que se utilizan las rutas locales en el sitio afectado.

Configuraciones de Oracle

Descripción general

El uso de la sincronización activa de SnapMirror no necesariamente agrega ni cambia ninguna práctica recomendada para operar una base de datos.

La mejor arquitectura depende de los requisitos del negocio. Por ejemplo, si el objetivo es contar con una protección RPO=0 frente a la pérdida de datos, pero el objetivo de tiempo de recuperación es relajado, el uso de bases de datos de instancia única de Oracle y la replicación de las LUN con SM-AS podrían ser suficientes y menos costoso desde la creación de licencias de Oracle. Un fallo del sitio remoto no interrumpiría las operaciones, y la pérdida del sitio principal provocaría que las LUN del sitio superviviente estén en línea y listas para utilizarse.

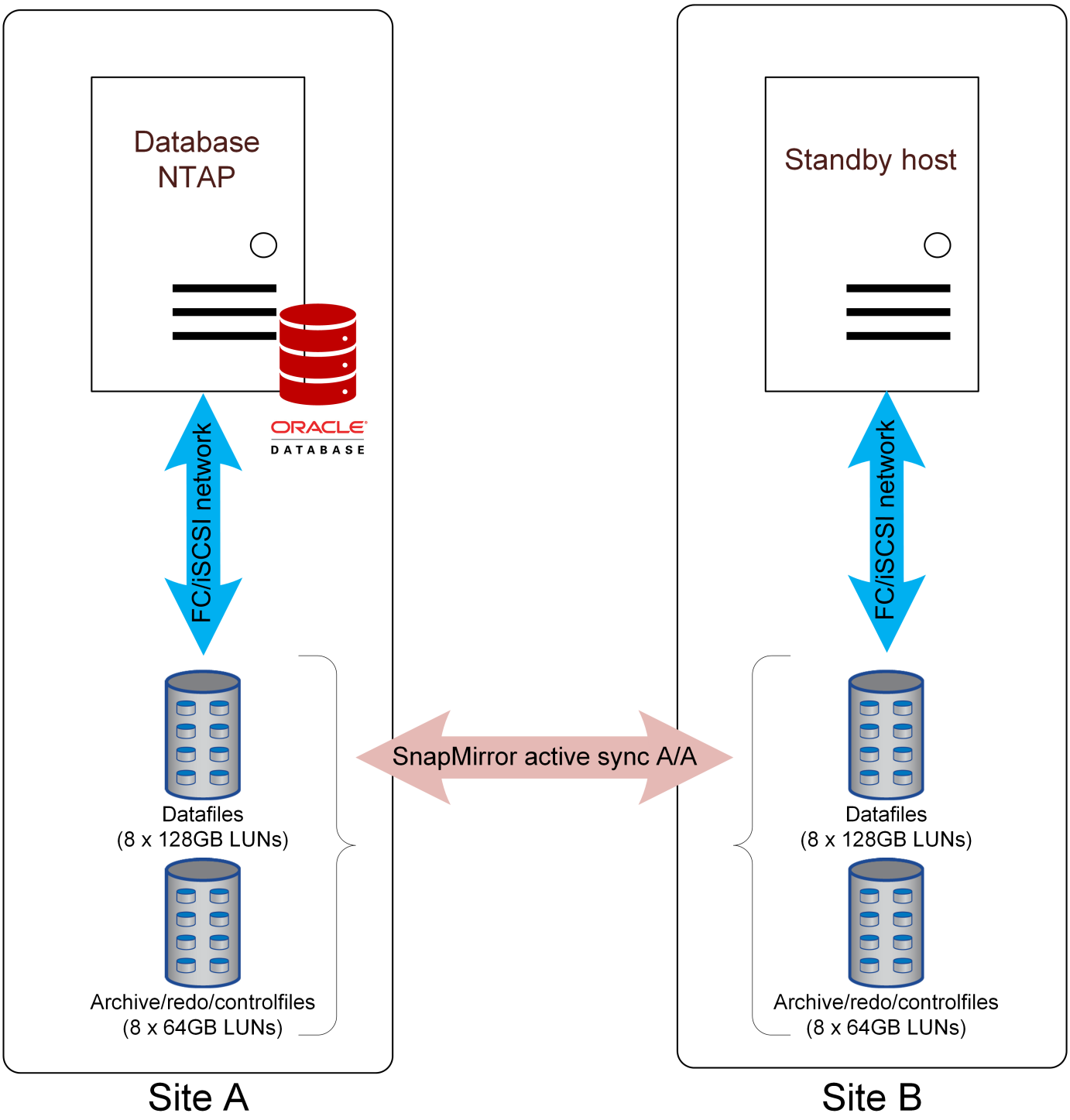
Si el objetivo de tiempo de recuperación fuera más estricto, la automatización activo-pasiva básica mediante scripts o clusterware como Pacemaker o Ansible mejoraría el tiempo de conmutación al nodo de respaldo. Por ejemplo, VMware HA podría configurarse para detectar un fallo de los equipos virtuales en el sitio primario y activar el equipo virtual en el sitio remoto.

Por último, para obtener una conmutación al respaldo extremadamente rápida, Oracle RAC se pudo poner en marcha en diferentes sitios. El RTO sería esencialmente cero porque la base de datos estaría en línea y disponible en ambas ubicaciones en todo momento.

Instancia única de Oracle

Los ejemplos que se explican a continuación muestran algunas de las muchas opciones para desplegar bases de datos de instancia única de Oracle con replicación de

sincronización activa de SnapMirror.



Conmutación al nodo de respaldo con un SO preconfigurado

La sincronización activa de SnapMirror ofrece una copia síncrona de los datos en el sitio de recuperación de desastres, pero para que los datos estén disponibles, es necesario 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 Pacemaker, se utilizan a menudo para crear un clúster en todos los sitios y, en muchos casos, el proceso de conmutación por error se puede ejecutar con scripts sencillos.

Si se pierden los nodos primarios, el clusterware (o scripts) pondrá las bases de datos en línea en el sitio alternativo. Una opción es crear servidores en espera preconfigurados para los recursos 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. Detecte el fallo del sitio principal
2. Realizar detección de LUN FC o iSCSI
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 `mount` un comando, y tanto las bases de datos como ASM se pueden iniciar y detener en la CLI con un único comando.

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. Detecte el fallo del sitio principal
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.

Protección frente a errores de almacenamiento

El diagrama anterior muestra el uso de "acceso no uniforme", donde la SAN no se extiende entre los sitios. Esto puede que sea más sencillo de configurar y, en algunos casos, puede que sea la única opción dadas las funcionalidades SAN actuales, pero también significa que un fallo del sistema de almacenamiento primario provocaría una interrupción en la base de datos hasta que se conmutara al nodo de respaldo de la aplicación.

Para una mayor resiliencia, la solución podría ponerse en marcha con "acceso uniforme". Esto permitiría a las aplicaciones seguir operando utilizando las rutas anunciadas desde el sitio opuesto.

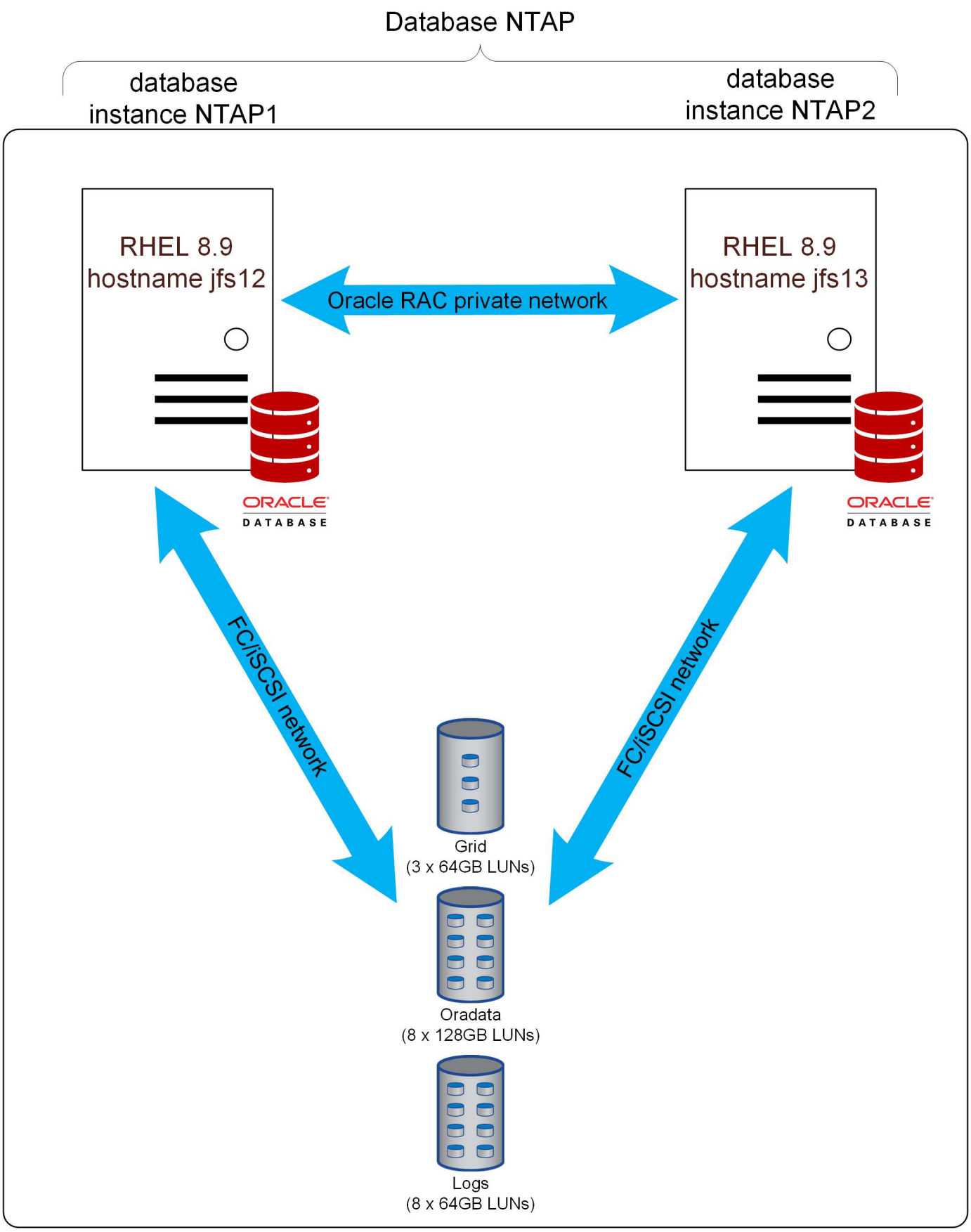
Oracle Extended RAC

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.

El cluster de RAC ampliado tradicional confiaba en la duplicación de ASM para proporcionar protección de datos. Este enfoque funciona, pero también requiere muchos pasos de configuración manuales e impone sobrecarga en la infraestructura de red. En cambio, permitir que SnapMirror Active Sync asuma la responsabilidad de la replicación de datos simplifica drásticamente la solución. Además, resulta más sencillo realizar operaciones como la sincronización, la resincronización después de interrupciones, las recuperaciones tras fallos y la gestión del quórum, además de que la SAN no tiene que distribuirse entre sitios, lo que simplifica el diseño y la gestión de SAN.

Replicación

Lo fundamental para comprender la funcionalidad de RAC en la sincronización activa de SnapMirror es ver el almacenamiento como un único conjunto de LUN alojadas en el almacenamiento reflejado. Por ejemplo:



No hay ninguna copia primaria ni copia duplicada. Lógicamente, solo existe una copia única de cada LUN, y esa LUN está disponible en rutas SAN ubicadas en dos sistemas de almacenamiento diferentes. Desde el punto de vista del host, no hay conmutación por error del almacenamiento; en cambio, existen cambios de

ruta. Varios eventos de fallo pueden provocar la pérdida de ciertas rutas a la LUN mientras otras rutas permanecen en línea. La sincronización activa de SnapMirror garantiza que los mismos datos estén disponibles en todas las rutas operativas.

Configuración del almacenamiento

En esta configuración de ejemplo, los discos ASM están configurados de la misma forma que en cualquier configuración de RAC de sitio único en el almacenamiento empresarial. Dado que el sistema de almacenamiento proporciona protección de datos, se utilizaría la redundancia externa de ASM.

Acceso uniforme frente a acceso no informado

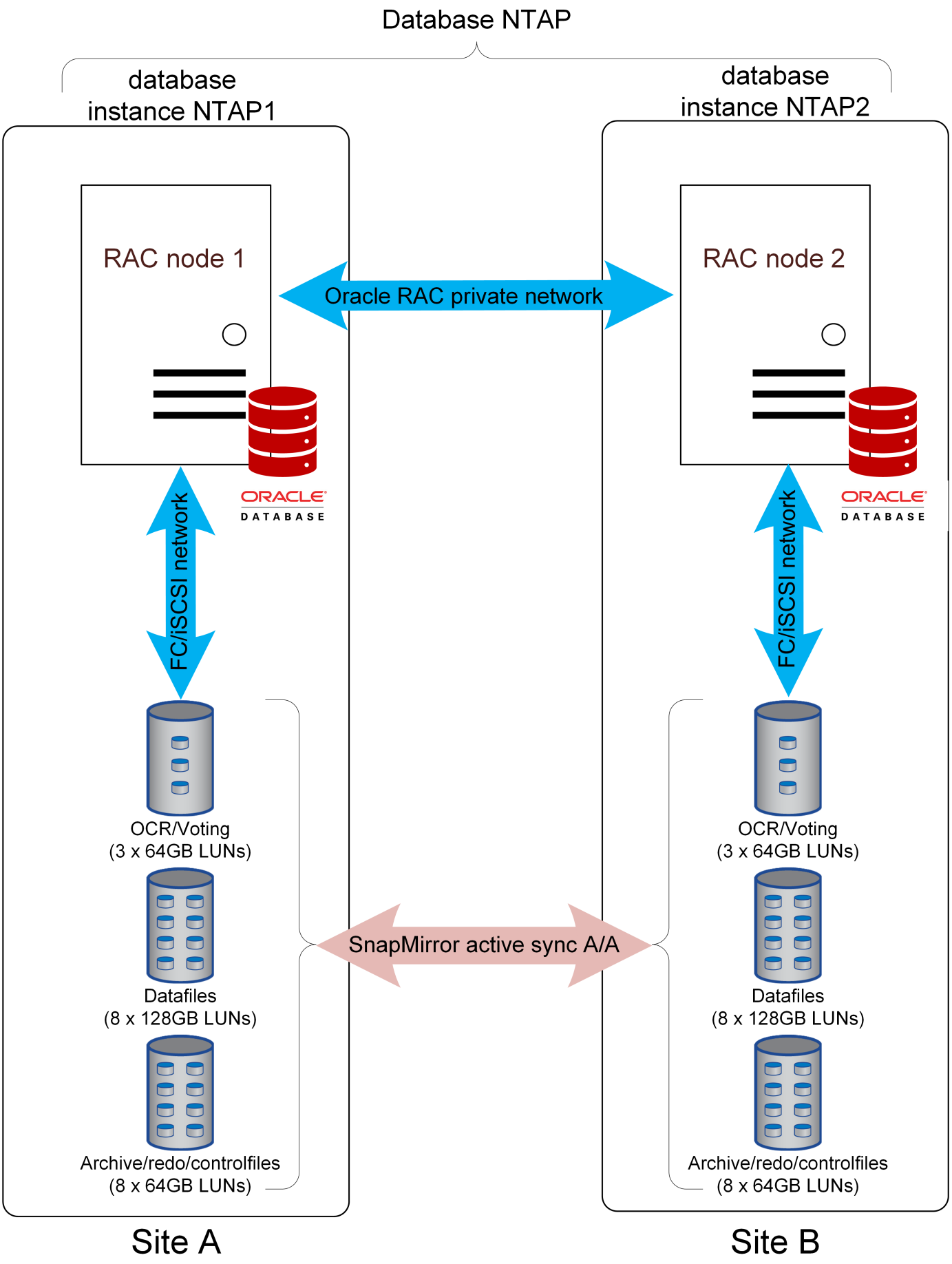
La consideración más importante con Oracle RAC en la sincronización activa de SnapMirror es si se debe utilizar un acceso uniforme o no uniforme.

El acceso uniforme significa que cada host puede ver rutas en ambos clústeres. El acceso no uniforme significa que los hosts solo pueden ver rutas al clúster local.

Ninguna de las opciones se recomienda o desaconseja específicamente. Algunos clientes disponen de fibra oscura disponible en todo momento para conectar sitios, mientras que otros no tienen esta conectividad o su infraestructura SAN no admite ISL de larga distancia.

Acceso no uniforme

El acceso no uniforme es más sencillo de configurar desde la perspectiva de SAN.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
SnapMirror active sync A/A
OCR/Voting (3 x 64GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

El principal "acceso no uniforme"inconveniente del método es que la pérdida de conectividad con ONTAP entre sitios o la pérdida de un sistema de almacenamiento supondrá la pérdida de instancias de base de datos en un sitio. Obviamente, esto no es deseable, pero puede ser un riesgo aceptable a cambio de una configuración SAN simple.

Acceso uniforme

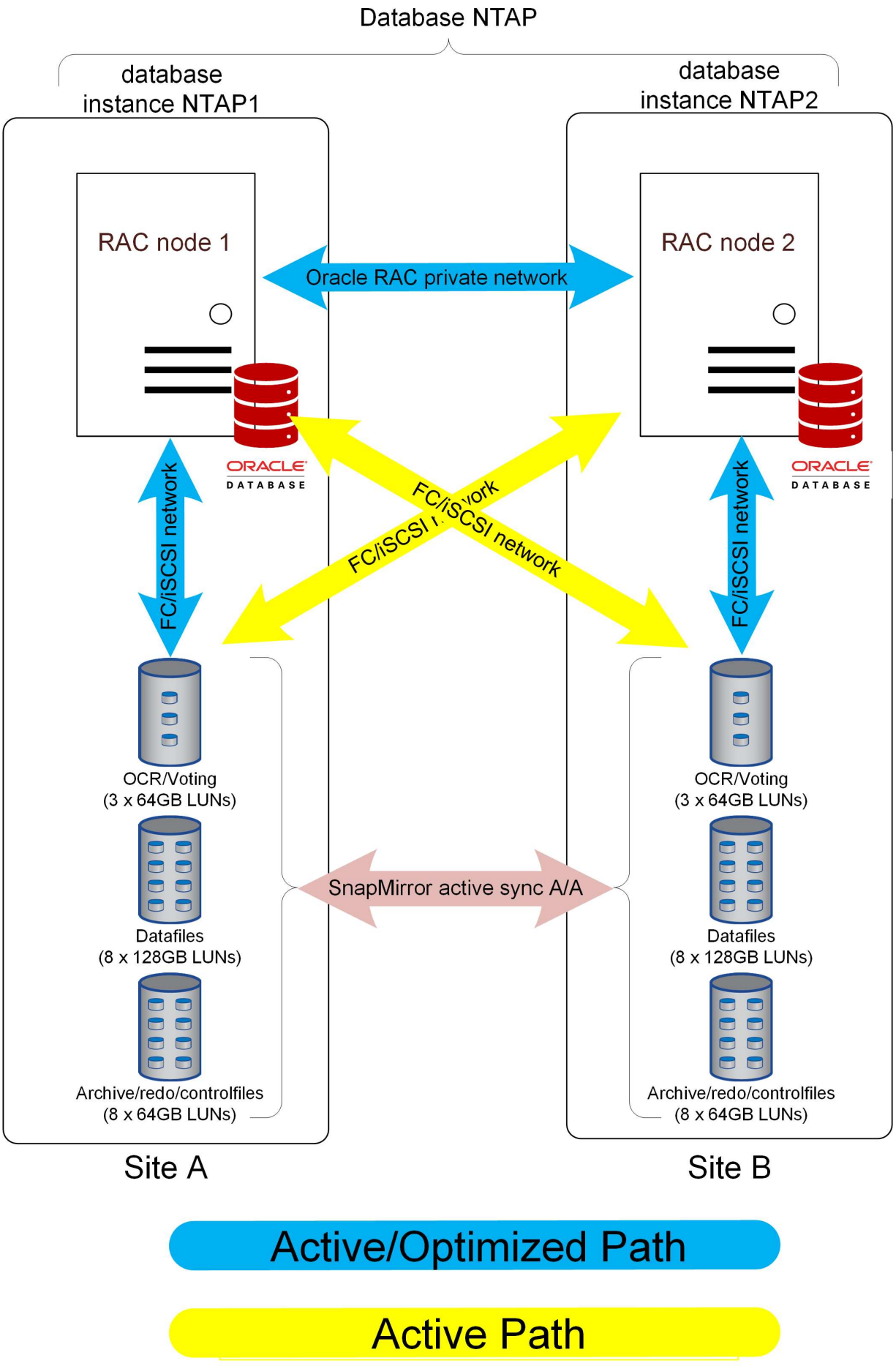
Para el acceso uniforme es necesario ampliar la SAN a través de varios sitios. La ventaja principal es que la pérdida de un sistema de almacenamiento no provocará la pérdida de una instancia de la base de datos. En su lugar, provocaría un cambio de multivía en el que se usan las rutas actualmente.

Hay varias formas de configurar el acceso no uniforme.

> En los diagramas que aparecen a continuación, también existen rutas activas pero no optimizadas que se utilizarán durante los simples fallos de controladoras, pero esas rutas no se muestran en interés de simplificar los diagramas.

AFF con ajustes de proximidad

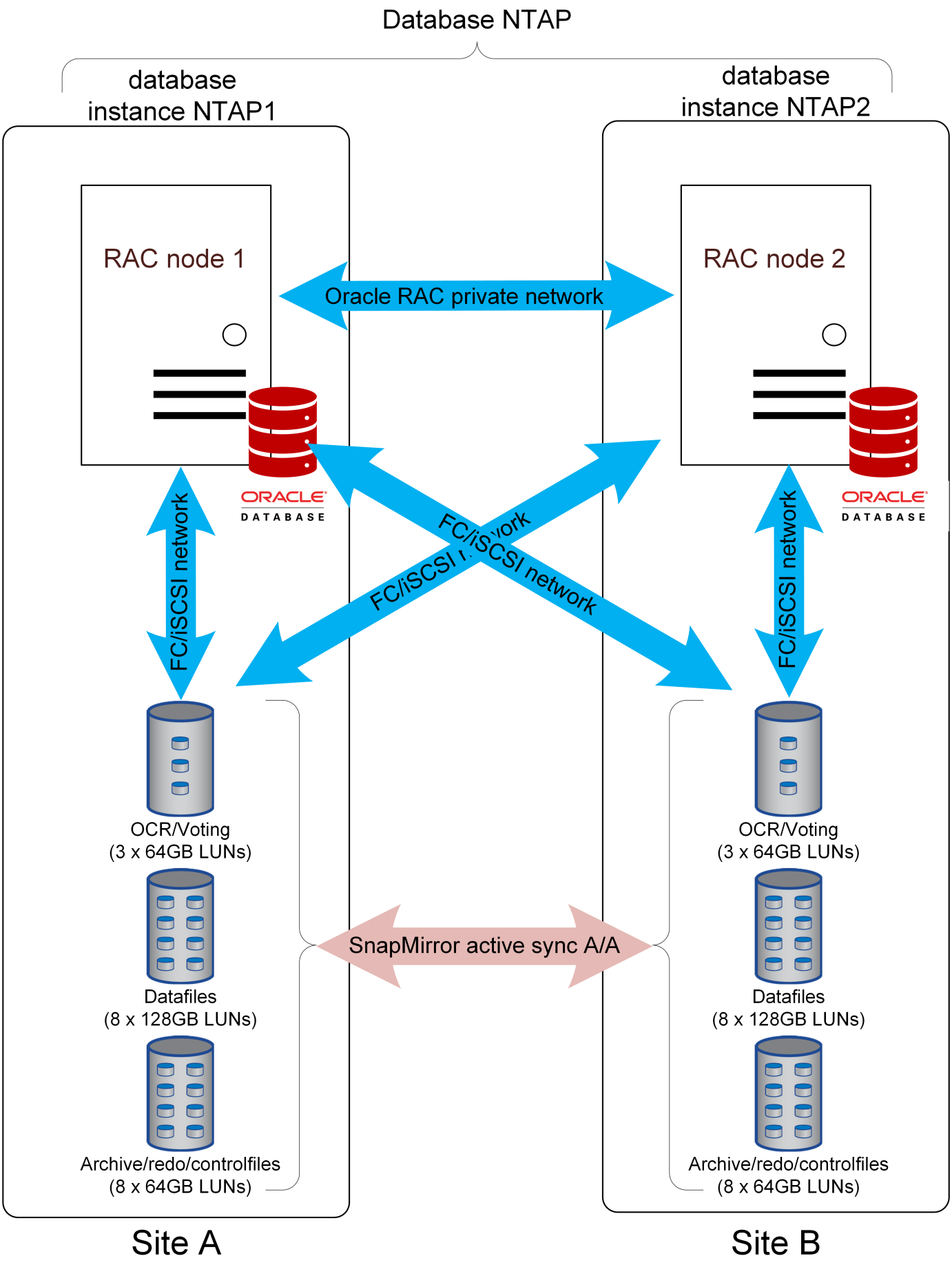
Si hay una latencia significativa entre los sitios, los sistemas AFF se pueden configurar con los ajustes de proximidad del host. Esto permite que cada sistema de almacenamiento tenga en cuenta qué hosts son locales y remotos y asigne prioridades de ruta según corresponda.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B
Active/Optimized Path
Active Path

En el funcionamiento normal, cada instancia de Oracle utilizaría preferentemente las rutas de acceso locales activas/optimizadas. El resultado es que la copia local de los bloques suministraría todas las lecturas. Esto proporciona la menor latencia posible. El I/O de escritura se envía de manera similar por rutas a la controladora local. La I/O debe replicarse antes de reconocerse y, por lo tanto, se produciría la latencia adicional al cruzar la red sitio a sitio, pero esto no se puede evitar en una solución de replicación síncrona.

ASA / AFF sin ajustes de proximidad

Si no hay latencia significativa entre los sitios, los sistemas AFF se pueden configurar sin ajustes de proximidad del host o se puede utilizar ASA.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

Cada host podrá utilizar todas las rutas operativas en ambos sistemas de almacenamiento. Esto mejora significativamente el rendimiento, ya que permite que cada host aproveche el potencial de rendimiento de dos clústeres, no solo uno.

Con ASA, no solo todas las rutas a ambos clústeres se considerarían activas y optimizadas, sino que las rutas en las controladoras de los partners también estarían activas. El resultado serían rutas SAN activas en todo el clúster, todo el tiempo.

> Los sistemas ASA también pueden usarse en una configuración de acceso no uniforme. Como no existen rutas entre sitios, no habría impacto en el rendimiento como resultado de I/O al cruzar el ISL.

RAC tiebreaker

Si bien el RAC extendido que usa SnapMirror active sync es una arquitectura simétrica con respecto a E/S, hay una excepción que está conectada a la gestión de cerebro dividido.

¿Qué sucede si el enlace de replicación se pierde y ninguno de los sitios tiene quórum? ¿Qué debería pasar? Esta pregunta se aplica tanto al comportamiento de Oracle RAC como al de ONTAP. Si los cambios no se pueden replicar en todos los sitios y desea reanudar las operaciones, uno de los sitios tendrá que sobrevivir y el otro sitio tendrá que dejar de estar disponible.

El "Mediador ONTAP" resuelve este requisito en la capa ONTAP. Hay varias opciones para RAC tiebreaking.

Oracle tiebreakers

El mejor método para gestionar los riesgos de Oracle RAC de cerebro dividido es utilizar un número impar de nodos de RAC, preferiblemente mediante el uso de un tiebreaker de sitio 3rd. Si un sitio 3rd no está disponible, la instancia de tiebreaker podría colocarse en un sitio de los dos sitios, designándolo efectivamente como un sitio de supervivencia preferido.

Oracle y css_critical

Con un número par de nodos, el comportamiento por defecto de Oracle RAC es que uno de los nodos del cluster se considerará más importante que el resto de nodos. El sitio con ese nodo de mayor prioridad sobrevivirá al aislamiento del sitio mientras que los nodos del otro sitio desalojarán. La priorización se basa en varios factores, pero también puede controlar este comportamiento mediante la `css_critical` configuración.

En la "ejemplo" arquitectura, los nombres de host de los nodos RAC son jfs12 y jfs13. Los ajustes actuales de `css_critical` son los siguientes:

```
[root@jfs12 ~]# /grid/bin/crsctl get server css_critical
CRS-5092: Current value of the server attribute CSS_CRITICAL is no.

[root@jfs13 trace]# /grid/bin/crsctl get server css_critical
CRS-5092: Current value of the server attribute CSS_CRITICAL is no.
```

Si desea que el sitio con jfs12 sea el sitio preferido, cambie este valor a yes en un sitio Un nodo y reinicie los servicios.

```
[root@jfs12 ~]# /grid/bin/crsctl set server css_critical yes
CRS-4416: Server attribute 'CSS_CRITICAL' successfully changed. Restart
Oracle High Availability Services for new value to take effect.

[root@jfs12 ~]# /grid/bin/crsctl stop crs
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'jfs12'
CRS-2673: Attempting to stop 'ora.crsd' on 'jfs12'
CRS-2790: Starting shutdown of Cluster Ready Services-managed resources on
server 'jfs12'
CRS-2673: Attempting to stop 'ora.ntap.ntappdb1.pdb' on 'jfs12'
…
CRS-2673: Attempting to stop 'ora.gipcd' on 'jfs12'
CRS-2677: Stop of 'ora.gipcd' on 'jfs12' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'jfs12' has completed
CRS-4133: Oracle High Availability Services has been stopped.

[root@jfs12 ~]# /grid/bin/crsctl start crs
CRS-4123: Oracle High Availability Services has been started.
```

Escenarios de fallo

Descripción general

La planificación de una arquitectura completa de aplicaciones de sincronización activa de SnapMirror requiere comprender cómo SM-AS responderá en varias situaciones de conmutación por error planificadas e imprevistas.

Para los siguientes ejemplos, supongamos que el sitio A está configurado como el sitio preferido.

Pérdida de conectividad de replicación

Si se interrumpe la replicación de SM-AS, la E/S de escritura no se puede completar porque sería imposible que un clúster replique los cambios en el sitio opuesto.

Sitio A (Sitio preferido)

El resultado de un fallo del enlace de replicación en el sitio preferido será una pausa de aproximadamente 15 segundos en el procesamiento de I/O de escritura, ya que ONTAP reintenta las operaciones de escritura replicadas antes de que determine que el enlace de replicación es realmente inaccesible. Una vez transcurridos los 15 segundos, el sistema del sitio A reanuda el procesamiento de E/S de lectura y escritura. Las rutas de SAN no se modificarán y los LUN permanecerán en línea.

Centro B

Dado que el sitio B no es el sitio preferido de sincronización activa de SnapMirror, sus rutas de LUN dejarán de estar disponibles después de unos 15 segundos.

Fallo del sistema de almacenamiento

El resultado de un fallo del sistema de almacenamiento es casi idéntico al de perder el enlace de replicación. El sitio superviviente debería experimentar una pausa de IO de aproximadamente 15 segundos. Una vez transcurrido ese período de 15 segundos, IO se reanudará en ese sitio como de costumbre.

Pérdida del mediador

El servicio de mediador no controla directamente las operaciones de almacenamiento. Funciona como una ruta de control alternativa entre los clústeres. Existe principalmente para automatizar la conmutación al nodo de respaldo sin el riesgo de un escenario de cerebro dividido. En un funcionamiento normal, cada clúster está replicando los cambios en su compañero y, por lo tanto, cada clúster puede verificar que el clúster asociado esté en línea y sirviendo datos. Si el enlace de replicación falla, la replicación se detendría.

El motivo por el que se necesita un mediador para una conmutación por error automatizada segura es que, de otro modo, sería imposible que un clúster de almacenamiento pueda determinar si la pérdida de comunicación bidireccional se debió a una interrupción de la red o a un error real de almacenamiento.

El mediador proporciona una ruta alternativa para que cada clúster compruebe el estado de su compañero. Los escenarios son los siguientes:

- Si un clúster puede ponerse en contacto directamente con su socio, los servicios de replicación están operativos. No se requiere ninguna acción.
- Si un sitio preferido no puede ponerse en contacto con su partner directamente o a través del mediador, se asumirá que el partner no está disponible o que se ha aislado y ha desconectado las rutas de LUN. El sitio preferido procederá a liberar el estado RPO=0 y continuará procesando las I/O de lectura y escritura.
- Si un sitio no preferido no puede ponerse en contacto directamente con su socio, pero puede contactarlo a través del mediador, tomará sus rutas fuera de línea y esperará la devolución de la conexión de replicación.
- Si un sitio no preferido no puede contactar a su partner directamente o a través de un mediador operativo, asumirá que el partner no está disponible o que se ha aislado y ha desconectado las rutas de LUN. El sitio no preferido continuará liberando el estado RPO=0 y continuará procesando las I/O de lectura y escritura. Asumirá el rol del origen de replicación y se convertirá en el nuevo sitio preferido.

Si el mediador no está totalmente disponible:

- El fallo en los servicios de replicación por cualquier motivo, incluido el fallo del sitio o del sistema de almacenamiento no preferido, provocará que el sitio preferido libere el estado RPO=0 y reanude el procesamiento de I/O de lectura y escritura. El sitio no preferido desconectará sus rutas.
- Un fallo del sitio preferido provocará una interrupción porque el sitio no preferido no podrá verificar que el sitio opuesto esté realmente fuera de línea y, por lo tanto, no sería seguro para el sitio no preferido reanudar los servicios.

Restauración de servicios

Tras resolver un fallo, como restaurar la conectividad de sitio a sitio o encender un sistema fallido, los extremos de sincronización activa de SnapMirror detectan automáticamente la presencia de una relación de replicación defectuosa y la devuelven a un estado RPO=0. Una vez que se restablece la replicación síncrona, las rutas fallidas volverán a conectarse.

En muchos casos, las aplicaciones en clúster detectan automáticamente el retorno de las rutas fallidas, y dichas aplicaciones también volverán a estar online. En otros casos, puede ser necesario un análisis SAN a nivel de host o es posible que las aplicaciones deban volver a conectarse manualmente. Depende de la aplicación y cómo se configura, y en general tales tareas se pueden automatizar fácilmente. El propio ONTAP

se repara automáticamente y no debería requerir la intervención del usuario para reanudar las operaciones de almacenamiento RPO=0.

Recuperación manual tras fallos

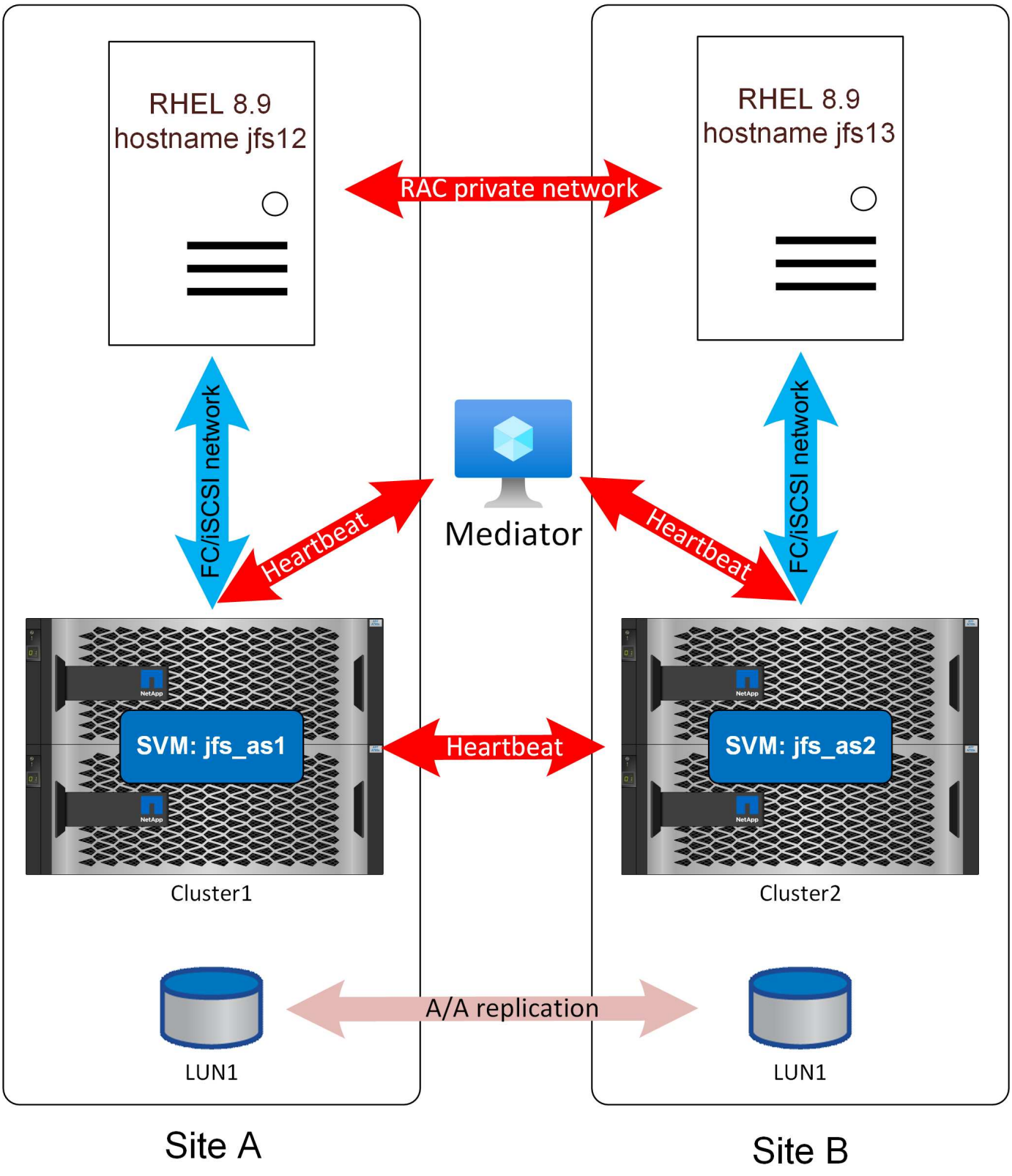
Cambiar el sitio preferido requiere una operación simple. I/O se detendrá durante un segundo o dos como autoridad sobre los cambios en el comportamiento de replicación entre los clústeres, pero I/O de otro modo no se verá afectado.

Arquitectura de ejemplo

Los ejemplos de fallos detallados que se muestran en estas secciones se basan en la arquitectura que se muestra a continuación.

> Esta es solo una de las muchas opciones para bases de datos Oracle en sincronización activa de SnapMirror. Este diseño fue elegido porque ilustra algunos de los escenarios más complicados.

En este diseño, asuma que el sitio A se establece en el "sitio preferido".



Fallo de interconexión de RAC

La pérdida del enlace de replicación de Oracle RAC producirá un resultado similar a la pérdida de conectividad de SnapMirror, excepto que los tiempos de espera serán más cortos por defecto. En la configuración predeterminada, un nodo de Oracle RAC

esperará 200 segundos después de la pérdida de conectividad de almacenamiento antes de expulsarlo, pero solo esperará 30 segundos después de la pérdida del latido de la red de RAC.

Los mensajes de CRS son similares a los que se muestran a continuación. Puede ver el lapso de tiempo de espera de 30 segundos. Dado que css_critical se estableció en jfs12, ubicado en el sitio A, ese será el sitio para sobrevivir y jfs13 en el sitio B será desalojado.

```
2024-09-12 10:56:44.047 [ONMD(3528)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 6.980 seconds
2024-09-12 10:56:48.048 [ONMD(3528)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 2.980 seconds
2024-09-12 10:56:51.031 [ONMD(3528)]CRS-1607: Node jfs13 is being evicted
in cluster incarnation 621599354; details at (:CSSNM00007:) in
/gridbase/diag/crs/jfs12/crs/trace/onmd.trc.
2024-09-12 10:56:52.390 [CRSD(6668)]CRS-7503: The Oracle Grid
Infrastructure process 'crsd' observed communication issues between node
'jfs12' and node 'jfs13', interface list of local node 'jfs12' is
'192.168.30.1:33194;', interface list of remote node 'jfs13' is
'192.168.30.2:33621;'.
2024-09-12 10:56:55.683 [ONMD(3528)]CRS-1601: CSSD Reconfiguration
complete. Active nodes are jfs12 .
2024-09-12 10:56:55.722 [CRSD(6668)]CRS-5504: Node down event reported for
node 'jfs13'.
2024-09-12 10:56:57.222 [CRSD(6668)]CRS-2773: Server 'jfs13' has been
removed from pool 'Generic'.
2024-09-12 10:56:57.224 [CRSD(6668)]CRS-2773: Server 'jfs13' has been
removed from pool 'ora.NTAP'.
```

Fallo de comunicación de SnapMirror

Si el enlace de replicación de sincronización activa de SnapMirror, el I/O de escritura no se puede completar porque sería imposible que un clúster replique los cambios en el sitio opuesto.

Centro a

El resultado en el sitio A de un fallo de enlace de replicación será una pausa de aproximadamente 15 segundos en el procesamiento de E/S de escritura, ya que ONTAP intenta replicar las escrituras antes de determinar que el enlace de replicación es realmente inoperable. Después de transcurridos 15 segundos, el clúster de ONTAP en el sitio A reanuda el procesamiento de I/O de lectura y escritura. Las rutas de SAN no se modificarán y los LUN permanecerán en línea.

Centro B

Dado que el sitio B no es el sitio preferido de sincronización activa de SnapMirror, sus rutas de LUN dejarán de estar disponibles después de unos 15 segundos.

El enlace de replicación se cortó en la marca de tiempo 15:19:44. La primera advertencia de Oracle RAC llega 100 segundos después cuando se acerca el timeout de 200 segundos (controlado por el parámetro de Oracle RAC disktimeout).

```
2024-09-10 15:21:24.702 [ONMD(2792)]CRS-1615: No I/O has completed after
50% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 99340 milliseconds.
2024-09-10 15:22:14.706 [ONMD(2792)]CRS-1614: No I/O has completed after
75% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 49330 milliseconds.
2024-09-10 15:22:44.708 [ONMD(2792)]CRS-1613: No I/O has completed after
90% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 19330 milliseconds.
2024-09-10 15:23:04.710 [ONMD(2792)]CRS-1604: CSSD voting file is offline:
/dev/mapper/grid2; details at (:CSSNM00058:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc.
2024-09-10 15:23:04.710 [ONMD(2792)]CRS-1606: The number of voting files
available, 0, is less than the minimum number of voting files required, 1,
resulting in CSSD termination to ensure data integrity; details at
(:CSSNM00018:) in /gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-10 15:23:04.716 [ONMD(2792)]CRS-1699: The CSS daemon is
terminating due to a fatal error from thread:
clssnmvDiskPingMonitorThread; Details at (:CSSSC00012:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-10 15:23:04.731 [OCSSD(2794)]CRS-1652: Starting clean up of CRSD
resources.
```

Una vez alcanzado el tiempo de espera del disco de quorum de 200 segundos, este nodo de Oracle RAC se expulsará del cluster y se reiniciará.

Fallo total de interconexión de red

Si el enlace de replicación entre las ubicaciones se pierde por completo, se interrumpirán tanto la conectividad de SnapMirror Active Sync como la de Oracle RAC.

La detección de cerebro dividido de Oracle RAC depende de los latidos del corazón del almacenamiento de Oracle RAC. Si la pérdida de conectividad de sitio a sitio provoca una pérdida simultánea de los latidos de la red de RAC y de los servicios de replicación de almacenamiento, el resultado es que los sitios de RAC no podrán comunicarse entre sitios ni a través de la interconexión de RAC ni de los discos de quorum de RAC. El resultado de un conjunto de nodos de numeración uniforme puede ser la expulsión de ambos sitios en la configuración predeterminada. El comportamiento exacto dependerá de la secuencia de eventos y del tiempo de la red RAC y de los sondeos de latidos del disco.

El riesgo de una interrupción del servicio de 2 sitios puede abordarse de dos maneras. En primer lugar, se

puede utilizar una "tiebreaker" configuración.

Si no hay un sitio 3rd disponible, este riesgo se puede solucionar ajustando el parámetro misscount en el cluster RAC. Bajo los valores predeterminados, el tiempo de espera de latido de la red RAC es de 30 segundos. RAC lo utiliza normalmente para identificar los nodos de RAC fallidos y quitarlos del cluster. También tiene una conexión con el latido del disco de votación.

Si, por ejemplo, un retroexcavador corta el conducto que transporta el tráfico entre sitios tanto para Oracle RAC como para los servicios de replicación de almacenamiento, comenzará la cuenta atrás de 30 segundos de recuento de errores. Si el nodo de sitio preferido de RAC no puede restablecer el contacto con el sitio opuesto en 30 segundos, y tampoco puede utilizar los discos de votación para confirmar que el sitio opuesto está caído dentro de la misma ventana de 30 segundos, entonces los nodos de sitio preferidos también expulsarán. El resultado es una interrupción completa de la base de datos.

Dependiendo de cuándo se produzca el sondeo de recuento incorrecto, es posible que 30 segundos no sean suficientes para que se agote el tiempo de espera de la sincronización activa de SnapMirror y que el almacenamiento del sitio preferido reanude los servicios antes de que caduque la ventana de 30 segundos. Esta ventana de 30 segundos se puede aumentar.

```
[root@jfs12 ~]# /grid/bin/crsctl set css misscount 100
CRS-4684: Successful set of parameter misscount to 100 for Cluster
Synchronization Services.
```

Este valor permite que el sistema de almacenamiento del sitio preferido reanude las operaciones antes de que se agote el tiempo de espera del recuento erróneo. A continuación, el resultado solo se expulsará de los nodos del sitio donde se quitaron las rutas de LUN. Ejemplo a continuación:

```
2024-09-12 09:50:59.352 [ONMD(681360)]CRS-1612: Network communication with
node jfs13 (2) has been missing for 50% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 49.570 seconds
2024-09-12 09:51:10.082 [CRSD(682669)]CRS-7503: The Oracle Grid
Infrastructure process 'crsd' observed communication issues between node
'jfs12' and node 'jfs13', interface list of local node 'jfs12' is
'192.168.30.1:46039;', interface list of remote node 'jfs13' is
'192.168.30.2:42037;'.
2024-09-12 09:51:24.356 [ONMD(681360)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 24.560 seconds
2024-09-12 09:51:39.359 [ONMD(681360)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 9.560 seconds
2024-09-12 09:51:47.527 [OHASD(680884)]CRS-8011: reboot advisory message
from host: jfs13, component: cssagent, with time stamp: L-2024-09-12-
09:51:47.451
2024-09-12 09:51:47.527 [OHASD(680884)]CRS-8013: reboot advisory message
text: oracssdagent is about to reboot this node due to unknown reason as
it did not receive local heartbeats for 10470 ms amount of time
2024-09-12 09:51:48.925 [ONMD(681360)]CRS-1632: Node jfs13 is being
removed from the cluster in cluster incarnation 621596607
```

Los Servicios de Soporte Oracle no recomiendan modificar los parámetros misscount o disktimeout para resolver problemas de configuración. Sin embargo, los cambios de estos parámetros pueden garantizarse e evitarse en muchos casos, incluido el arranque SAN, la virtualización y las configuraciones de replicación del almacenamiento. Si, por ejemplo, tenía problemas de estabilidad con una red SAN o IP que provocaba expulsiones de RAC, debería corregir el problema subyacente y no cargar los valores del recuento de errores o el tiempo de espera del disco. Cambiar los tiempos de espera para corregir los errores de configuración es enmascarar un problema, no resolver un problema. Cambiar estos parámetros para configurar correctamente un entorno RAC basado en aspectos de diseño de la infraestructura subyacente es diferente y es coherente con las sentencias de soporte de Oracle. Con el arranque SAN, es común ajustar misscount hasta 200 para que coincida con el tiempo de espera del disco. Consulte "este enlace" para obtener información adicional.

Error en el centro

El resultado de un fallo del sistema de almacenamiento o del sitio es casi idéntico al resultado de perder el enlace de replicación. El sitio superviviente debería experimentar una pausa de I/O de aproximadamente 15 segundos en las escrituras. Una vez transcurrido ese período de 15 segundos, IO se reanudará en ese sitio como de costumbre.

Si solo el sistema de almacenamiento se vio afectado, el nodo de Oracle RAC del sitio donde se ha producido el fallo perderá los servicios de almacenamiento e introducirá la misma cuenta atrás con un tiempo de espera de disco de 200 segundos antes de su expulsión y reinicio posterior.

```
2024-09-11 13:44:38.613 [ONMD(3629)]CRS-1615: No I/O has completed after
50% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 99750 milliseconds.
2024-09-11 13:44:51.202 [ORAAGENT(5437)]CRS-5011: Check of resource "NTAP"
failed: details at "(:CLSN00007:)" in
"/gridbase/diag/crs/jfs13/crs/trace/crsd_oraagent_oracle.trc"
2024-09-11 13:44:51.798 [ORAAGENT(75914)]CRS-8500: Oracle Clusterware
ORAAGENT process is starting with operating system process ID 75914
2024-09-11 13:45:28.626 [ONMD(3629)]CRS-1614: No I/O has completed after
75% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 49730 milliseconds.
2024-09-11 13:45:33.339 [ORAAGENT(76328)]CRS-8500: Oracle Clusterware
ORAAGENT process is starting with operating system process ID 76328
2024-09-11 13:45:58.629 [ONMD(3629)]CRS-1613: No I/O has completed after
90% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 19730 milliseconds.
2024-09-11 13:46:18.630 [ONMD(3629)]CRS-1604: CSSD voting file is offline:
/dev/mapper/grid2; details at (:CSSNM00058:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc.
2024-09-11 13:46:18.631 [ONMD(3629)]CRS-1606: The number of voting files
available, 0, is less than the minimum number of voting files required, 1,
resulting in CSSD termination to ensure data integrity; details at
(:CSSNM00018:) in /gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-11 13:46:18.638 [ONMD(3629)]CRS-1699: The CSS daemon is
terminating due to a fatal error from thread:
clssnmvDiskPingMonitorThread; Details at (:CSSSC00012:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-11 13:46:18.651 [OCSSD(3631)]CRS-1652: Starting clean up of CRSD
resources.
```

El estado de la ruta de SAN en el nodo RAC que ha perdido los servicios de almacenamiento se parece a este:

```
oradata7 (3600a0980383041334a3f55676c697347) dm-20 NETAPP,LUN C-Mode
size=128G features='3 queue_if_no_path pg_init_retries 50' hwhandler='1
alua' wp=rw
|-+- policy='service-time 0' prio=0 status=enabled
| `- 34:0:0:18 sdam 66:96  failed faulty running
`-+- policy='service-time 0' prio=0 status=enabled
  `- 33:0:0:18 sdaj 66:48  failed faulty running
```

El host linux detectó la pérdida de las rutas mucho más rápido que 200 segundos, pero desde el punto de vista de la base de datos, las conexiones del cliente al host en el sitio con errores se seguirán congelando durante 200 segundos en la configuración predeterminada de Oracle RAC. Las operaciones de base de datos completa solo se reanudarán una vez que se complete el expulsión.

Mientras tanto, el nodo de Oracle RAC en la ubicación opuesta registrará la pérdida del otro nodo de RAC. De lo contrario, sigue funcionando como de costumbre.

```
2024-09-11 13:46:34.152 [ONMD(3547)]CRS-1612: Network communication with
node jfs13 (2) has been missing for 50% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 14.020 seconds
2024-09-11 13:46:41.154 [ONMD(3547)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 7.010 seconds
2024-09-11 13:46:46.155 [ONMD(3547)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 2.010 seconds
2024-09-11 13:46:46.470 [OHASD(1705)]CRS-8011: reboot advisory message
from host: jfs13, component: cssmonit, with time stamp: L-2024-09-11-
13:46:46.404
2024-09-11 13:46:46.471 [OHASD(1705)]CRS-8013: reboot advisory message
text: At this point node has lost voting file majority access and
oracssdmonitor is rebooting the node due to unknown reason as it did not
receive local hearbeats for 28180 ms amount of time
2024-09-11 13:46:48.173 [ONMD(3547)]CRS-1632: Node jfs13 is being removed
from the cluster in cluster incarnation 621516934
```

Fallo del mediador

El servicio de mediador no controla directamente las operaciones de almacenamiento. Funciona como una ruta de control alternativa entre los clústeres. Existe principalmente para automatizar la conmutación al nodo de respaldo sin el riesgo de un escenario de cerebro dividido.

En un funcionamiento normal, cada clúster está replicando los cambios en su compañero y, por lo tanto, cada clúster puede verificar que el clúster asociado esté en línea y sirviendo datos. Si el enlace de replicación falla, la replicación se detendría.

El motivo por el que se necesita un mediador para llevar a cabo operaciones automatizadas seguras es que, de otro modo, sería imposible que los clústeres de almacenamiento puedan determinar si la pérdida de comunicación bidireccional se debió a una interrupción de la red o a un error real de almacenamiento.

El mediador proporciona una ruta alternativa para que cada clúster compruebe el estado de su compañero. Los escenarios son los siguientes:

- Si un clúster puede ponerse en contacto directamente con su socio, los servicios de replicación están operativos. No se requiere ninguna acción.
- Si un sitio preferido no puede ponerse en contacto con su partner directamente o a través del mediador, se asumirá que el partner no está disponible o que se ha aislado y ha desconectado las rutas de LUN. El sitio preferido procederá a liberar el estado RPO=0 y continuará procesando las I/O de lectura y escritura.
- Si un sitio no preferido no puede ponerse en contacto directamente con su socio, pero puede contactarlo a través del mediador, tomará sus rutas fuera de línea y esperará la devolución de la conexión de replicación.

- Si un sitio no preferido no puede contactar a su partner directamente o a través de un mediador operativo, asumirá que el partner no está disponible o que se ha aislado y ha desconectado las rutas de LUN. El sitio no preferido continuará liberando el estado RPO=0 y continuará procesando las I/O de lectura y escritura. Asumirá el rol del origen de replicación y se convertirá en el nuevo sitio preferido.

Si el mediador no está totalmente disponible:

- Un fallo en los servicios de replicación por cualquier motivo provocará que el sitio preferido publique el estado RPO=0 y reanude el procesamiento de I/O de lectura y escritura. El sitio no preferido desconectará sus rutas.
- Un fallo del sitio preferido provocará una interrupción porque el sitio no preferido no podrá verificar que el sitio opuesto esté realmente fuera de línea y, por lo tanto, no sería seguro para el sitio no preferido reanudar los servicios.

Restauración del servicio

SnapMirror es reparación automática. La sincronización activa de SnapMirror detecta automáticamente la presencia de una relación de replicación defectuosa y la devuelve a un estado RPO=0. Una vez que se restablece la replicación síncrona, las rutas volverán a conectarse.

En muchos casos, las aplicaciones en clúster detectan automáticamente el retorno de las rutas fallidas, y dichas aplicaciones también volverán a estar online. En otros casos, puede ser necesario un análisis SAN a nivel de host o es posible que las aplicaciones deban volver a conectarse manualmente.

Depende de la aplicación y de cómo se configura, y en general tales tareas se pueden automatizar fácilmente. La sincronización activa de SnapMirror es autocorregida y no debería requerir la intervención del usuario para reanudar las operaciones de almacenamiento RPO=0 una vez que se restablezcan la alimentación y la conectividad.

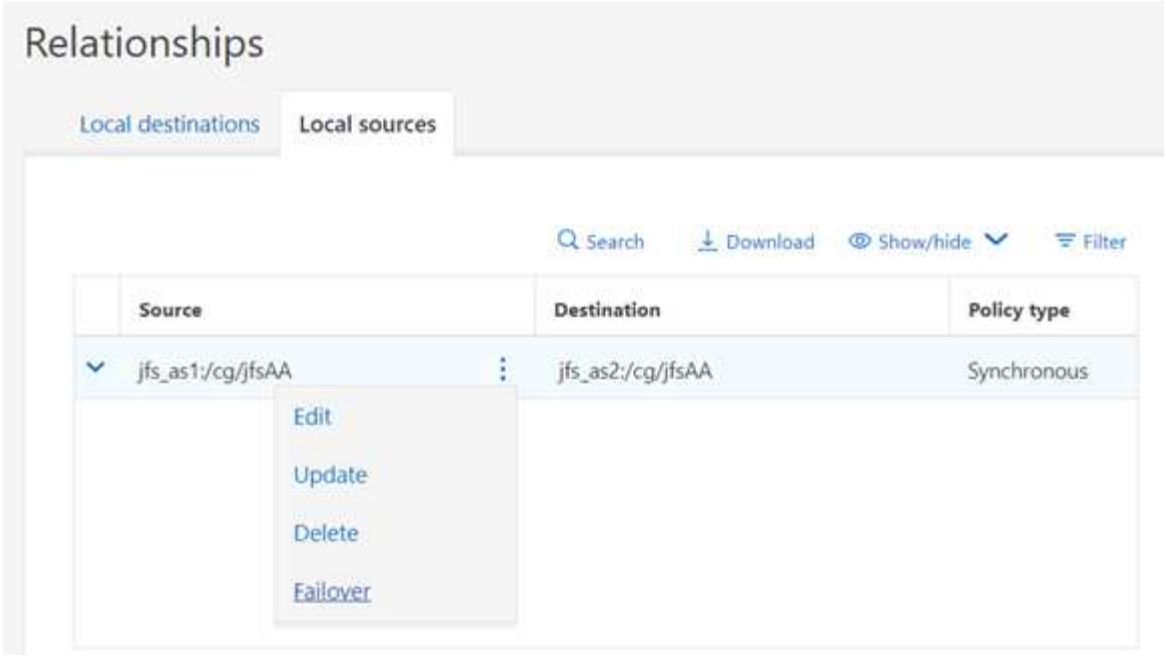
Recuperación manual tras fallos

El término «conmutación por error» no hace referencia a la dirección de la replicación con el servicio de sincronización activa de SnapMirror porque es una tecnología de replicación bidireccional. En su lugar, la recuperación tras fallos hace referencia al sistema de almacenamiento en el sitio preferido en caso de fallo.

Por ejemplo, puede que desee realizar una conmutación al respaldo para cambiar el sitio preferido antes de apagar un sitio por mantenimiento o antes de realizar una prueba de recuperación ante desastres.

Cambiar el sitio preferido requiere una operación simple. I/O se detendrá durante un segundo o dos como autoridad sobre los cambios en el comportamiento de replicación entre los clústeres, pero I/O de otro modo no se verá afectado.

Ejemplo de interfaz gráfica de usuario:



Ejemplo de cambio a través de la CLI:

```
Cluster2::> snapmirror failover start -destination-path jfs_as2:/cg/jfsAA
[Job 9575] Job is queued: SnapMirror failover for destination
"jfs_as2:/cg/jfsAA                      ".

Cluster2::> snapmirror failover show

Source    Destination                                           Error
Path      Path        Type     Status    start-time end-time   Reason
--------- ----------  -------- --------- ---------- ---------- ----------
jfs_as1:/cg/jfsAA
          jfs_as2:/cg/jfsAA
                      planned  completed 9/11/2024  9/11/2024
                                         09:29:22   09:29:32

The new destination path can be verified as follows:


Cluster1::> snapmirror show -destination-path jfs_as1:/cg/jfsAA

                           Source Path: jfs_as2:/cg/jfsAA
                        Destination Path: jfs_as1:/cg/jfsAA
                       Relationship Type: XDP
                 Relationship Group Type: consistencygroup
                  SnapMirror Policy Type: automated-failover-duplex
                       SnapMirror Policy: AutomatedFailOverDuplex
                             Tries Limit: -
                           Mirror State: Snapmirrored
                    Relationship Status: InSync
```