Base de datos Oracle

Enterprise applications

NetApp
January 12, 2026

Tabla de contenidos

Base de datos Oracle

Bases de datos de Oracle en ONTAP

ONTAP está diseñado para bases de datos de Oracle. Durante décadas, ONTAP se ha optimizado para las demandas específicas de las I/O de las bases de datos relacionales y se crearon varias funciones de ONTAP específicamente para satisfacer las necesidades de las bases de datos de Oracle e incluso a petición de la misma Oracle Inc.

> Esta documentación sustituye a los informes técnicos *TR-3633 publicados anteriormente: Bases de datos Oracle en ONTAP; TR-4591: Protección de datos de Oracle: Backup, recuperación, replicación; TR-4592: Oracle en MetroCluster; y TR-4534: Migración de bases de datos de Oracle a sistemas de almacenamiento de NetApp*

Además de las muchas formas posibles en que ONTAP aporta valor a su entorno de bases de datos, también presenta una amplia variedad de requisitos de usuario, como el tamaño de la base de datos, los requisitos de rendimiento y las necesidades de protección de datos. Las puestas en marcha conocidas del almacenamiento de NetApp incluyen todo, desde un entorno virtualizado de aproximadamente 6.000 bases de datos que se ejecutan con VMware ESX hasta un almacén de datos de instancia única con un tamaño actualmente de 996TB TB, que sigue creciendo. Como resultado, existen pocas mejores prácticas claras para configurar una base de datos Oracle en un almacenamiento de NetApp.

Los requisitos para operar una base de datos Oracle en el almacenamiento de NetApp se tratan de dos formas. En primer lugar, cuando existe una práctica recomendada clara, se llamará específicamente. En un nivel general, se explicarán muchas consideraciones de diseño que deben tratar los arquitectos de las soluciones de almacenamiento de Oracle basadas en sus requisitos empresariales específicos.

Configuración de ONTAP en sistemas AFF/ FAS

RAID

RAID se refiere al uso de redundancia para proteger los datos contra la pérdida de una unidad.

De vez en cuando se plantean preguntas sobre los niveles de RAID en la configuración del almacenamiento NetApp utilizado para las bases de datos de Oracle y otras aplicaciones empresariales. Muchas de las mejores prácticas de Oracle heredadas sobre la configuración de la cabina de almacenamiento contienen advertencias sobre el uso de mirroring de RAID y/o la prevención de ciertos tipos de RAID. Aunque plantean puntos válidos, estas fuentes no se aplican a RAID 4 y a las tecnologías de NetApp RAID DP y RAID-TEC utilizadas en ONTAP.

RAID 4, RAID 5, RAID 6, RAID DP y RAID-TEC usan la paridad para garantizar que el fallo de una unidad no provoque la pérdida de datos. Estas opciones de RAID ofrecen un aprovechamiento del almacenamiento mucho mejor en comparación con mirroring, pero la mayoría de las implementaciones de RAID tienen un inconveniente que afecta a las operaciones de escritura. La finalización de una operación de escritura en otras implementaciones de RAID puede requerir varias lecturas de unidad para volver a generar los datos de paridad, un proceso comúnmente denominado penalización de RAID.

Sin embargo, ONTAP no implica este proceso de penalización por RAID. Esto se debe a la integración de

NetApp WAFL (Write Anywhere File Layout) con la capa RAID. Las operaciones de escritura se fusionan en la RAM y se preparan como una franja RAID completa, incluida la generación de paridad. ONTAP no necesita realizar una lectura para completar una escritura, lo que significa que ONTAP y WAFL evitan la penalización de RAID. El rendimiento de las operaciones cruciales para la latencia, como el registro de reconstrucción, no se ve afectado, y las escrituras de archivos de datos aleatorios no suponen ningún tipo de penalización de RAID por la necesidad de regenerar la paridad.

En cuanto a la fiabilidad estadística, incluso RAID DP ofrece una mejor protección que el mirroring RAID. El problema principal es la demanda que se realiza en las unidades durante una recompilación de RAID. Con un conjunto RAID reflejado, el riesgo de que se pierdan datos tras el fallo en una unidad durante la reconstrucción a su compañero en el conjunto RAID es mucho mayor que el riesgo de un fallo de triple unidad en un conjunto RAID DP.

Gestión de la capacidad

Gestionar una base de datos u otra aplicación empresarial con almacenamiento empresarial predecible, gestionable y de alto rendimiento requiere cierto espacio libre en las unidades para la gestión de datos y metadatos. La cantidad de espacio libre necesario depende del tipo de unidad utilizada y los procesos empresariales.

El espacio libre se define como el espacio que no se usa para datos reales e incluye espacio sin asignar en el propio agregado y el espacio no utilizado dentro de los volúmenes constituyentes. También se debe tener en cuenta el thin provisioning. Por ejemplo, un volumen puede contener 1TB 000 LUN de las cuales solo el 50% es utilizado por datos reales. En un entorno con thin provisioning, parece que esto consume 500GB TB de espacio de manera correcta. Sin embargo, en un entorno totalmente aprovisionado, parece que toda la capacidad de 1TB está en uso. Los 500GB GB de espacio no asignado están ocultos. Los datos reales no utilizan este espacio y, por lo tanto, debe incluirse en el cálculo del espacio libre total.

Las recomendaciones de NetApp para los sistemas de almacenamiento que se utilizan para aplicaciones empresariales son las siguientes:

Agregados SSD, incluidos los sistemas AFF

NetApp recomienda un mínimo de 10% de espacio libre. Esto incluye todo el espacio no utilizado, incluido el espacio libre dentro del agregado o un volumen y cualquier espacio libre que se asigne debido al uso de aprovisionamiento completo, pero que los datos reales no usan. El espacio lógico no es importante, la pregunta es cuánto espacio físico libre real está disponible para el almacenamiento de datos.

La recomendación de un 10% de espacio libre es muy conservadora. Los agregados SSD pueden admitir cargas de trabajo con niveles de utilización aún mayores sin afectar en absoluto al rendimiento. No obstante, a medida que aumenta el uso del agregado, también aumenta el riesgo de quedarse sin espacio si no se supervisa el uso de forma cuidadosa. Además, aunque ejecutar un sistema a un 99 % de capacidad puede que no afecte al rendimiento, probablemente se traduciría en un esfuerzo de gestión al intentar evitar que se llene completamente mientras se solicita hardware adicional. Además, la adquisición e instalación de unidades adicionales puede demorar algún tiempo.

Agregados de HDD, incluidos los agregados de Flash Pool

NetApp recomienda un mínimo de 15% de espacio libre cuando se utilizan unidades giratorias. Esto incluye todo el espacio no utilizado, incluido el espacio libre dentro del agregado o un volumen y cualquier espacio libre que se asigne debido al uso de aprovisionamiento completo, pero que los datos reales no usan. El rendimiento se verá afectado cuando el espacio libre se acerque al 10%.

Máquinas virtuales de almacenamiento

La gestión del almacenamiento de bases de datos de Oracle se centraliza en una máquina virtual de almacenamiento (SVM).

Una SVM, conocida como Vserver en la interfaz de línea de comandos de ONTAP, es una unidad funcional básica de almacenamiento, lo que resulta útil comparar una SVM con una máquina virtual «guest» en un servidor VMware ESX.

Cuando se instala por primera vez, ESX no tiene capacidades preconfiguradas, como alojar un sistema operativo invitado o admitir una aplicación de usuario final. Es un contenedor vacío hasta que se define una máquina virtual (VM). ONTAP es similar. Cuando ONTAP se instala por primera vez, no cuenta con funcionalidades de servicio de datos hasta que se crea una SVM. Es la personalidad de la SVM que define los servicios de datos.

Al igual que otros aspectos de la arquitectura de almacenamiento, las mejores opciones para el diseño de SVM y de la interfaz lógica (LIF) dependen en gran medida de los requisitos de escalado y las necesidades del negocio.

SVM

No existe ninguna práctica recomendada oficial para el aprovisionamiento de SVM para ONTAP. El método correcto depende de los requisitos de gestión y seguridad.

La mayoría de los clientes operan un SVM principal para la mayoría de sus requisitos diarios y después crean un número pequeño de SVM para necesidades especiales. Por ejemplo, es posible que desee crear:

- SVM para una base de datos empresarial crítica gestionada por un equipo especializado
- Una SVM para un grupo de desarrollo al que se le ha otorgado un control administrativo completo para que pueda gestionar su propio almacenamiento de forma independiente
- Una máquina virtual de almacenamiento SVM para datos empresariales confidenciales, como datos de recursos humanos o informes financieros, a los que debe limitarse el equipo administrativo

En un entorno multi-tenant, los datos de cada inquilino pueden recibir una SVM dedicada. El límite del número de SVM y LIF por clúster, pareja de alta disponibilidad y nodo dependen del protocolo que se utilice, del modelo de nodo y de la versión de ONTAP. Consulte la "Hardware Universe de NetApp" para estos límites.

Gestión del rendimiento con QoS de ONTAP

La gestión segura y eficaz de varias bases de datos Oracle requiere una estrategia de QoS eficaz. La razón es el aumento constante en las funcionalidades de rendimiento de un sistema de almacenamiento moderno.

En concreto, la creciente adopción del almacenamiento all-flash ha permitido consolidar las cargas de trabajo. Las cabinas de almacenamiento que se basan en medios giratorios tendían a admitir solo una cantidad limitada de cargas de trabajo con un gran volumen de I/O debido a las funcionalidades de IOPS limitadas de la

tecnología de unidades rotacionales más antigua. Una o dos bases de datos altamente activas saturarían las unidades subyacentes mucho antes de que las controladoras de almacenamiento alcanzaran sus límites. Esto ha cambiado. La funcionalidad de rendimiento de un número relativamente pequeño de unidades SSD puede saturar incluso las controladoras de almacenamiento más potentes. Esto significa que pueden aprovecharse todas las funcionalidades de las controladoras sin miedo al colapso repentino del rendimiento cuando se disparan los picos de latencia de los medios giratorios.

Como ejemplo de referencia, un sencillo sistema AFF A800 de alta disponibilidad de dos nodos es capaz de dar servicio a hasta un millón de IOPS aleatorias antes de que la latencia aumente por encima del milisegundo. Sería de esperar que muy pocas cargas de trabajo individuales alcancen estos niveles. El uso completo de esta cabina para el sistema A800 de AFF implicará alojar múltiples cargas de trabajo y hacerlo de forma segura y, al mismo tiempo, garantizar la previsibilidad, requiere controles de calidad de servicio.

Existen dos tipos de calidad de servicio en ONTAP: IOPS y ancho de banda. Los controles de calidad de servicio se pueden aplicar a SVM, volúmenes, LUN y archivos.

Calidad de servicio IOPS

Obviamente, un control de calidad de servicio de IOPS se basa en el número total de IOPS de un recurso determinado, pero hay una serie de aspectos de la calidad de servicio de IOPS que quizá no sean intuitivos. Al principio, algunos clientes se quedaron desconcertados por el aparente aumento de la latencia cuando se alcanza un umbral de IOPS. El aumento de la latencia es el resultado natural de la limitación de IOPS. Lógicamente, funciona de forma similar a un sistema de tokens. Por ejemplo, si un volumen determinado que contiene archivos de datos tiene un límite de 10K IOPS, cada I/O que llegue primero deberá recibir un token para continuar con el procesamiento. Mientras no se hayan consumido más de 10K tokens en un segundo determinado, no hay retrasos. Si las operaciones de I/O deben esperar para recibir el token, esta espera aparece como latencia adicional. Cuanto más fuerte sea una carga de trabajo que supere el límite de calidad de servicio, más tiempo debe esperar cada I/O en la cola para su procesamiento, lo cual parece que el usuario tiene una mayor latencia.

Tenga cuidado al aplicar controles QoS a los datos de transacción/redo log de la base de datos. Si bien las demandas de rendimiento del redo log suelen ser mucho más bajas que las de los archivos de datos, la actividad de redo log es rápida. El E/S se produce en pulsos breves y un límite de QoS que parece adecuado para los niveles medios de E/S de redo puede ser demasiado bajo para los requisitos reales. El resultado puede ser limitaciones de rendimiento graves ya que QoS se conecta con cada ráfaga de redo log. En general, el redo y el registro de archivos no deben estar limitados por QoS.

Calidad del ancho de banda

No todos los tamaños de I/O son iguales. Por ejemplo, una base de datos puede estar realizando un gran número de lecturas de bloque pequeño, lo que haría que se alcance el umbral de IOPS. pero las bases de datos también pueden estar realizando una operación de exploración de tabla completa que consistiría en un número muy pequeño de lecturas de bloque grandes, lo que consume una gran cantidad de ancho de banda pero relativamente pocas IOPS.

Del mismo modo, un entorno VMware podría generar un gran número de IOPS aleatorias durante el arranque, pero realizaría menos I/O, pero más grandes, durante un backup externo.

A veces, para gestionar el rendimiento de forma efectiva se requieren límites de IOPS o de calidad de servicio del ancho de banda o incluso ambos.

Calidad de servicio mínima/garantizada

Muchos clientes buscan una solución que incluya una calidad de servicio garantizada, una solución que se pueda conseguir más de lo que parece y que potencialmente supone un derroche. Por ejemplo, colocar 10 bases de datos con una garantía de 10K IOPS requiere configurar un sistema para un escenario en el que las 10 bases de datos se ejecuten simultáneamente a 10K 000 IOPS, para un total de 100K 000.

El mejor uso para los controles mínimos de calidad de servicio es proteger las cargas de trabajo cruciales. Por ejemplo, piense en una controladora ONTAP con un número máximo de IOPS de 500K KB posible y una combinación de cargas de trabajo de producción y desarrollo. Debe aplicar políticas de calidad de servicio máximas a las cargas de trabajo de desarrollo para evitar que una base de datos determinada monopolice la controladora. A continuación, aplicaría políticas mínimas de calidad de servicio a las cargas de trabajo de producción para asegurarse de que siempre tengan las IOPS necesarias disponibles cuando las necesite.

Calidad de servicio adaptativa

La calidad de servicio adaptativa se refiere a la función ONTAP, donde el límite de calidad de servicio se basa en la capacidad del objeto de almacenamiento. Rara vez se utiliza con bases de datos porque normalmente no hay ningún vínculo entre el tamaño de una base de datos y sus requisitos de rendimiento. Las bases de datos de gran tamaño pueden ser casi inertes, mientras que las bases de datos más pequeñas pueden ser las más intensivas en IOPS.

La calidad de servicio adaptativa puede resultar muy útil con los almacenes de datos de virtualización porque los requisitos de IOPS de dichos conjuntos de datos tienden a correlacionarse con el tamaño total de la base de datos. Es probable que los almacenes de datos más recientes que contienen 1TB TB de archivos VMDK requieran la mitad de rendimiento que un almacén de datos de 2TB GB. La calidad de servicio adaptativa le permite aumentar automáticamente los límites de calidad de servicio a medida que el almacén de datos se llena con datos.

Eficiencia

Las funciones de gestión eficiente del espacio de ONTAP se optimizan para las bases de datos de Oracle. En casi todos los casos, el mejor método es dejar los valores predeterminados con todas las funciones de eficiencia activadas.

Las funciones de eficiencia del espacio, como la compresión, la compactación y la deduplicación están diseñadas para aumentar la cantidad de datos lógicos que se adaptan a una determinada cantidad de almacenamiento físico. El resultado es una reducción de los costes y los gastos generales de gestión.

En un nivel superior, la compresión es un proceso matemático por el cual los patrones en los datos se detectan y codifican de manera que reducen los requisitos de espacio. Por el contrario, la deduplicación detecta bloques de datos repetidos y elimina las copias externas. La compactación permite que varios bloques lógicos de datos compartan el mismo bloque físico en medios.

Consulte las siguientes secciones sobre thin provisioning para obtener una explicación de la interacción entre la eficiencia del almacenamiento y la reserva fraccionaria.

Compresión

Antes de la disponibilidad de sistemas de almacenamiento all-flash, la compresión basada en cabinas era de un valor limitado debido a que la mayoría de las cargas de trabajo con un gran volumen de I/O requerían un gran número de discos para proporcionar un rendimiento aceptable. Los sistemas de almacenamiento contenían invariablemente mucha más capacidad de la necesaria como efecto secundario al gran número de

unidades. La situación ha cambiado con el aumento del almacenamiento de estado sólido. Ya no es necesario sobreaprovisionar enormemente las unidades solo para obtener un buen rendimiento. El espacio de las unidades de un sistema de almacenamiento puede coincidir con las necesidades de capacidad reales.

La mayor funcionalidad de IOPS de las unidades de estado sólido (SSD) casi siempre genera ahorro de costes en comparación con las unidades giratorias, pero la compresión puede conseguir un mayor ahorro al aumentar la capacidad efectiva de los medios de estado sólido.

Existen varias formas de comprimir datos. Muchas bases de datos incluyen sus propias funcionalidades de compresión, pero esto se observa muy rara vez en los entornos del cliente. La razón suele ser la penalización de rendimiento para un **cambio** a los datos comprimidos, además con algunas aplicaciones hay altos costos de licencia para la compresión a nivel de base de datos. Por último, existen las consecuencias de rendimiento generales para las operaciones de base de datos. Tiene poco sentido pagar un alto coste de licencia por CPU por una CPU que realiza compresión y descompresión de datos en lugar de trabajo real de base de datos. Una mejor opción es descargar el trabajo de compresión en el sistema de almacenamiento.

Compresión adaptativa

La compresión adaptativa se ha probado minuciosamente en cargas de trabajo empresariales sin que ello afecte al rendimiento, incluso en un entorno all-flash en el que la latencia se mide en microsegundos. Algunos clientes incluso han informado de un aumento del rendimiento con el uso de la compresión, ya que los datos siguen comprimidos en la caché, lo que aumenta efectivamente la cantidad de caché disponible en una controladora.

ONTAP gestiona bloques físicos en 4KB unidades. La compresión adaptativa usa un tamaño de bloque de compresión predeterminado de 8KB KB, lo que significa que los datos se comprimen en 8KB unidades. Esto coincide con el tamaño de bloque de 8KB KB que suelen utilizar las bases de datos relacionales. Los algoritmos de compresión son más eficientes a medida que se comprimen más datos como una sola unidad. Un tamaño de bloque de compresión de 32KB KB haría más eficiente el espacio que una unidad de bloques de compresión de 8KB KB. Esto significa que la compresión adaptativa con el tamaño de bloque de 8KB KB predeterminado conduce a tasas de eficiencia ligeramente más bajas, pero también ofrece una ventaja significativa si se usa un tamaño de bloque de compresión más pequeño. Las cargas de trabajo de bases de datos incluyen una gran cantidad de actividad de sobrescritura. Para sobrescribir un bloque de datos de 8KB GB de 32KB comprimido, es necesario volver a leer los 32KB TB completos de datos lógicos, descomprimirlos, actualizar la región de 8KB requerida, recomprimir y, a continuación, volver a escribir todo el 32KB en las unidades. Esta es una operación muy cara para un sistema de almacenamiento y es el motivo por el que algunas cabinas de almacenamiento de la competencia basadas en bloques de compresión más grandes también incurren en un impacto significativo en el rendimiento con las cargas de trabajo de base de datos.

> El tamaño de los bloques utilizado por la compresión adaptativa se puede aumentar hasta 32KB KB. Esto puede mejorar la eficiencia del almacenamiento y debe considerarse en el caso de archivos inactivos, como registros de transacciones y archivos de backup, cuando se almacena una cantidad sustancial de dichos datos en la cabina. En algunas situaciones, las bases de datos activas que usan un tamaño de bloque de 16KB KB o de 32KB KB también pueden beneficiarse de aumentar el tamaño de bloque de la compresión adaptativa para que coincida. Consulte a un representante de NetApp o de su partner para obtener orientación sobre si esto es adecuado para su carga de trabajo.

Los bloques de compresión superiores a los 8KB MB no se deben usar junto a la deduplicación en destinos de backup en streaming. El motivo es que los pequeños cambios en los datos de backup afectan a la ventana de compresión de 32KB:1. Si la ventana cambia, los datos comprimidos resultantes difieren en todo el archivo. La deduplicación ocurre después de la compresión, lo que significa que el motor de deduplicación ve cada backup comprimido de forma diferente. Si se requiere la deduplicación de backups en streaming, solo deberá usarse la compresión adaptativa de 8KB bloques. Es preferible recurrir a la compresión adaptativa, ya que funciona con un tamaño de bloque más pequeño y no interrumpe la eficiencia de la deduplicación. Por motivos similares, la compresión en el lado del host también interfiere con la eficiencia de la deduplicación.

Alineación de la compresión

La compresión adaptativa en un entorno de base de datos requiere tener en cuenta algún tipo de aspecto en la alineación de bloques de compresión. Hacerlo solo es una preocupación para los datos sujetos a sobrescrituras aleatorias de bloques muy específicos. Este enfoque es similar en concepto a la alineación general del sistema de archivos, donde el inicio de un sistema de archivos debe alinearse con un límite de dispositivo 4K y el tamaño de bloque de un sistema de archivos debe ser un múltiplo de 4K.

Por ejemplo, una escritura 8KB en un archivo se comprime solo si se alinea con un límite de 8KB KB en el propio sistema de archivos. Este punto significa que debe caer en los primeros 8KB del archivo, el segundo 8KB del archivo, y así sucesivamente. La forma más sencilla de garantizar una alineación correcta es utilizar el tipo de LUN correcto, cualquier partición creada debe tener un desplazamiento desde el inicio del dispositivo que sea un múltiplo de 8K y usar un tamaño de bloque del sistema de archivos que sea un múltiplo del tamaño del bloque de la base de datos.

Los datos como los backups o los registros de transacciones son operaciones escritas secuencialmente que abarcan varios bloques, todos ellos comprimidos. Por lo tanto, no hay necesidad de considerar la alineación. El único patrón de E/S preocupante es la sobrescritura aleatoria de archivos.

Compactación de datos

La compactación de datos es una tecnología que mejora la eficiencia de la compresión. Como se ha indicado anteriormente, la compresión adaptativa por sí sola puede proporcionar un ahorro de 2:1 KB, ya que se limita a almacenar una I/O de 8KB KB en un bloque de 4KB WAFL. Los métodos de compresión con tamaños de bloque más grandes ofrecen una mejor eficiencia. Sin embargo, no son adecuados para datos sujetos a sobrescrituras de bloques pequeños. La descompresión de 32KB unidades de datos, la actualización de una parte de 8KB, la recompresión y la escritura en las unidades genera una sobrecarga.

La compactación de datos permite almacenar varios bloques lógicos en bloques físicos. Por ejemplo, una base de datos con datos altamente comprimibles, como texto o bloques parcialmente completos, puede comprimirse de 8KB a 1KB. Sin compactación, esos 1KB TB de datos seguirían ocupando un bloque completo de 4KB KB. La compactación de datos inline permite almacenar 1KB TB de datos comprimidos en solo 1KB GB de espacio físico junto con otros datos comprimidos. No es una tecnología de compresión; simplemente es una forma más eficaz de asignar espacio en las unidades y, por tanto, no debe crear un efecto de rendimiento detectable.

El grado de ahorro obtenido varía. Por lo general, los datos que ya están comprimidos o cifrados no se pueden comprimir aún más y, por lo tanto, estos conjuntos de datos no se benefician de la compactación. Por el contrario, los archivos de datos recién inicializados que contienen poco más que metadatos de bloques y ceros se comprimen hasta 80:1.

Eficiencia de almacenamiento sensible a la temperatura

La eficiencia de almacenamiento sensible a la temperatura (TSSE) está disponible en ONTAP 9, e.8 y posteriores. Se basa en mapas de calor de acceso a bloques para identificar los bloques a los que se accede con poca frecuencia y comprimirlos con mayor eficiencia.

Deduplicación

La deduplicación es eliminar los tamaños de bloques duplicados de un conjunto de datos. Por ejemplo, si existiera el mismo bloque de 4KB KB en 10 archivos diferentes, la deduplicación redirigiría ese bloque de 4KB KB en los 10 archivos al mismo bloque físico de 4KB KB. El resultado sería una mejora de 10:1 veces en eficiencia en esos datos.

Los datos, como las LUN de arranque invitado de VMware, suelen deduplicar muy bien porque constan de varias copias de los mismos archivos del sistema operativo. Se ha observado una eficiencia de 100:1 y mayor.

Algunos datos no contienen datos duplicados. Por ejemplo, un bloque de Oracle contiene una cabecera que es única globalmente para la base de datos y un cola que es casi único. Como resultado, la deduplicación de una base de datos de Oracle rara vez produce un ahorro superior al 1%. La deduplicación con bases de datos de MS SQL es ligeramente mejor, pero los metadatos únicos a nivel de bloque siguen siendo una limitación.

En pocos casos, se ha observado un ahorro de espacio de hasta un 15 % en bases de datos con 16KB KB y tamaños de bloque grandes. El primer 4KB de cada bloque contiene el encabezado único a nivel mundial, y el último bloque de 4KB contiene el remolque casi único. Los bloques internos pueden optar a la deduplicación, aunque en la práctica esto se atribuye casi por completo a la deduplicación de datos puestos a cero.

Muchas cabinas de la competencia afirman la capacidad de deduplicar bases de datos basándose en la presunción de que una base de datos se copia varias veces. En este sentido, la deduplicación de NetApp también podría utilizarse, pero ONTAP ofrece una opción mejor: La tecnología FlexClone de NetApp. El resultado final es el mismo; se crean varias copias de una base de datos que comparten la mayoría de los bloques físicos subyacentes. El uso de FlexClone es mucho más eficiente que tomarse tiempo para copiar archivos de base de datos y después deduplicarlos. Es, de hecho, la no duplicación en lugar de la deduplicación, porque nunca se crea un duplicado.

Eficiencia y thin provisioning

Las funciones de eficiencia son formas de thin provisioning. Por ejemplo, una LUN de 100GB GB que ocupa un volumen de 100GB GB podría comprimirse hasta 50GB 000. Todavía no hay ahorros reales realizados porque el volumen sigue siendo de 100GB GB. Primero se debe reducir el volumen para que el espacio ahorrado se pueda usar en cualquier otro lugar del sistema. Si los cambios realizados en la LUN de 100GB TB más adelante hacen que los datos se puedan comprimir menos, el tamaño de la LUN aumentará y el volumen podría llenarse.

Se recomienda encarecidamente el aprovisionamiento ligero porque puede simplificar la gestión y, al mismo tiempo, proporcionar una mejora considerable en la capacidad utilizable con un ahorro de costes asociado. La razón es simple: Los entornos de bases de datos suelen incluir una gran cantidad de espacio vacío, un gran número de volúmenes y LUN, y datos comprimibles. El aprovisionamiento grueso provoca la reserva de espacio en el almacenamiento para volúmenes y LUN por si en algún momento llegan a estar llenos un 100 % y contienen un 100 % de datos que no se pueden comprimir. Es poco probable que esto ocurra. El thin provisioning permite reclamar y utilizar ese espacio en otra parte, y permite que la gestión de la capacidad se base en el propio sistema de almacenamiento en lugar de muchos volúmenes y LUN más pequeños.

Algunos clientes prefieren utilizar el aprovisionamiento pesado, ya sea para cargas de trabajo específicas o, por lo general, basándose en prácticas operativas y de adquisición establecidas.

Si un volumen se aprovisiona en exceso, debe tenerse cuidado desactivar por completo todas las funciones de eficiencia de ese volumen, incluida la descompresión y la eliminación de la deduplicación con `sis undo` el comando. El volumen no debe aparecer en `volume efficiency show` la salida. Si lo hace, el volumen sigue estando parcialmente configurado para las funciones de eficiencia. Como resultado, la sobrescritura garantiza un funcionamiento diferente, lo que aumenta la posibilidad de que las sobretensiones de la configuración hagan que el volumen se quede sin espacio inesperadamente, lo que producirá errores de I/O de la base de datos.

Mejores prácticas de eficiencia

NetApp recomienda lo siguiente:

Valores predeterminados de AFF

Los volúmenes creados en ONTAP en un sistema AFF all-flash son thin provisioning, con todas las funciones de eficiencia inline habilitadas. Aunque por lo general, las bases de datos no se benefician de la deduplicación y pueden incluir datos que no se pueden comprimir, la configuración predeterminada es adecuada para casi todas las cargas de trabajo. ONTAP está diseñado para procesar eficientemente todo tipo de datos y patrones de I/O, independientemente de que generen o no ahorros. Los valores predeterminados solo se deben cambiar si los motivos se entienden por completo y existe un beneficio para desviarse.

Recomendaciones generales

- Si los volúmenes o LUN no son con thin provisioning, debe deshabilitar todas las configuraciones de eficiencia, ya que el uso de estas funciones no proporciona ahorro y la combinación de aprovisionamiento grueso con la eficiencia de espacio habilitada puede provocar un comportamiento inesperado, incluidos errores de falta de espacio.
- Si los datos no están sujetos a sobrescrituras, como con backups o registros de transacciones de base de datos, puede lograr una mayor eficiencia habilitando TSSE con un bajo período de enfriamiento.
- Es posible que algunos archivos contengan una cantidad significativa de datos que no se puedan comprimir, por ejemplo, cuando la compresión ya está activada en el nivel de aplicación de los archivos está cifrada. Si se da alguna de estas situaciones, considere la posibilidad de deshabilitar la compresión para permitir un funcionamiento más eficiente en otros volúmenes que contengan datos comprimibles.
- No utilice la compresión 32KB ni la deduplicación con backups de bases de datos. Consulte la sección Compresión adaptativa para obtener más detalles.

Aprovisionamiento ligero

El thin provisioning para una base de datos de Oracle requiere una planificación cuidadosa porque el resultado es configurar más espacio en un sistema de almacenamiento del que necesariamente está disponible físicamente. Vale mucho la pena el esfuerzo porque, cuando se hace correctamente, el resultado es un ahorro significativo de costes y mejoras en la capacidad de gestión.

El thin provisioning se presenta de muchas formas y forma parte de muchas de las funciones que ofrece ONTAP para un entorno de aplicaciones empresariales. Además, thin provisioning está estrechamente relacionado con las tecnologías de eficiencia por el mismo motivo: Las funciones de eficiencia permiten almacenar más datos lógicos de lo que existen técnicamente en el sistema de almacenamiento.

Casi cualquier uso de las copias Snapshot implica thin provisioning. Por ejemplo, una base de datos de 10TB

TB típica en almacenamiento de NetApp incluye unos 30 días de copias Snapshot. Este arreglo da como resultado aproximadamente 10TB TB de datos visibles en el sistema de archivos activo y 300TB TB dedicados a las copias snapshot. El total de 310TB TB de almacenamiento suele residir en aproximadamente 12TB a 15TB GB de espacio. La base de datos activa consume 10TB GB y los 300TB TB restantes solo requieren de 2TB a 5TB GB de espacio, ya que solo se almacenan los cambios realizados en los datos originales.

La clonación es también un ejemplo de aprovisionamiento ligero. Un importante cliente de NetApp creó 40 clones de una base de datos de 80TB para que los utilizara el equipo de desarrollo. Si los 40 desarrolladores que utilizan estos clones sobrescribieran cada bloque en cada archivo de datos, se necesitarían más de 3,2PB GB de almacenamiento. En la práctica, la rotación es baja y el requisito de espacio colectivo se acerca a 40TB, ya que solo se almacenan cambios en las unidades.

Gestión del espacio

Se debe tener cierta precaución con el thin provisioning de un entorno de aplicaciones porque las tasas de cambios de los datos pueden aumentar de forma inesperada. Por ejemplo, el consumo de espacio debido a las instantáneas puede aumentar rápidamente si se reindexan las tablas de la base de datos o si se aplican parches a gran escala a los huéspedes de VMware. Una copia de seguridad fuera de lugar puede escribir una gran cantidad de datos en muy poco tiempo. Por último, puede ser difícil recuperar algunas aplicaciones si un sistema de archivos se queda sin espacio libre inesperadamente.

Afortunadamente, estos riesgos se pueden abordar con una cuidadosa configuración de `volume-autogrow` y.. `snapshot-autodelete` normativas. Como sus nombres implican, estas opciones permiten al usuario crear políticas que desactiven automáticamente el espacio consumido por las copias Snapshot o aumentar un volumen para alojar datos adicionales. Hay muchas opciones disponibles y las necesidades varían según el cliente.

Consulte "documentación de gestión de almacenamiento lógico" para obtener un análisis completo de estas funciones.

Reservas fraccionarias

La reserva fraccionaria es el comportamiento de una LUN en un volumen con respecto a la eficiencia del espacio. Cuando la opción `fractional-reserve` se establece en 100 %, todos los datos del volumen pueden experimentar una rotación del 100 % con cualquier patrón de datos sin agotar el espacio en el volumen.

Por ejemplo, piense en una base de datos en un único LUN de 250GB GB en un volumen de 1TB GB. La creación de una instantánea provocaría de inmediato la reserva de 250GB GB de espacio adicional en el volumen para garantizar que el volumen no se quede sin espacio por ningún motivo. El uso de reservas fraccionarias suele ser un desperdicio debido a que es extremadamente poco probable que cada byte del volumen de base de datos deba sobrescribirse. No hay razón para reservar espacio para un evento que nunca ocurre. Sin embargo, si un cliente no puede supervisar el consumo de espacio en un sistema de almacenamiento y debe tener la seguridad de que nunca se agota el espacio, se necesitarían reservas fraccionarias del 100% para utilizar copias Snapshot.

Compresión y deduplicación

La compresión y la deduplicación son ambas formas de thin provisioning. Por ejemplo, una huella de datos de 50TB MB puede comprimirse hasta 30TB MB, lo que supone un ahorro de 20TB MB. Para que la compresión proporcione beneficios, algunos de esos 20TB MB deben utilizarse para otros datos o el sistema de almacenamiento debe adquirirse con menos de 50TB TB. El resultado es almacenar más datos de los que están disponibles técnicamente en el sistema de almacenamiento. Desde el punto de vista de los datos, hay 50TB GB de datos, a pesar de que ocupa solo 30TB GB en las unidades.

Siempre existe la posibilidad de que cambie la capacidad de compresión de un conjunto de datos, lo que provocaría un aumento del consumo de espacio real. Este aumento del consumo significa que la compresión debe gestionarse como sucede con otras formas de thin provisioning en términos de supervisión y uso `volume-autogrow` y.. `snapshot-autodelete`.

La compresión y la deduplicación se tratan de forma más detallada en el enlace de sección:efficiency.html

Compresión y reservas fraccionarias

La compresión es una forma de thin provisioning. Las reservas fraccionarias afectan al uso de la compresión, con una nota importante; se reserva espacio con antelación para la creación de la instantánea. Normalmente, la reserva fraccionaria sólo es importante si existe una instantánea. Si no hay ninguna instantánea, la reserva fraccionaria no es importante. Este no es el caso con la compresión. Si se crea una LUN en un volumen con compresión, ONTAP conserva el espacio para acomodar una copia de Snapshot. Este comportamiento puede ser confuso durante la configuración, pero es esperado.

Como ejemplo, piense en un volumen de 10GB GB con una LUN de 5GB TB que se ha comprimido en 2,5GB sin copias Snapshot. Considere estos dos escenarios:

- La reserva fraccionaria = 100 da como resultado el uso de 7,5GB
- La reserva fraccionaria = 0 da como resultado el uso de 2,5GB

El primer escenario incluye 2,5GB GB de consumo de espacio para los datos actuales y 5GB GB de espacio para representar una rotación del 100% de la fuente antes del uso de la tecnología Snapshot. El segundo escenario no reserva espacio extra.

Aunque esta situación pueda parecer confusa, es poco probable que se encuentre en la práctica. La compresión implica thin provisioning y thin provisioning de un entorno de LUN requiere reservas fraccionarias. Siempre es posible que los datos comprimidos se sobrescriban en algo que no se pueda comprimir, lo que significa que un volumen debe estar aplicado mediante thin provisioning para que la compresión produzca ahorro.

NetApp recomienda las siguientes configuraciones de reserva:

- Configurado `fractional-reserve` a 0 cuando se implementa la supervisión de la capacidad básica junto con `volume-autogrow` y.. `snapshot-autodelete`.
- Configurado `fractional-reserve` a 100 si no hay capacidad de monitoreo o si es imposible agotar el espacio bajo cualquier circunstancia.

Espacio libre y asignación de espacio LVM

La eficiencia del aprovisionamiento fino de LUN activos en un entorno de sistema de archivos puede perderse con el tiempo a medida que se eliminan datos. A menos que los datos eliminados se sobrescriban con ceros (consulte también el enlace:oracle-storage-san-config-asm ru.html[ASMRU]) o se libere el espacio con la recuperación de espacio TRIM/UNMAP, los datos "borrados" ocupan cada vez más espacio en blanco no asignado en el sistema de archivos. Además, el aprovisionamiento fino de LUN activos tiene una utilidad limitada en muchos entornos de bases de datos porque los archivos de datos se inicializan a su tamaño completo en el momento de su creación.

Una planificación cuidadosa de la configuración de LVM puede mejorar la eficiencia y minimizar la necesidad de aprovisionar el almacenamiento y redimensionar las LUN. Cuando se utiliza un LVM como Veritas VxVM u Oracle ASM, los LUN subyacentes se dividen en extensiones que solo se utilizan cuando es necesario. Por ejemplo, si un conjunto de datos empieza con un tamaño de 2TB GB, pero podría crecer hasta 10TB TB con

el tiempo, este conjunto de datos podría colocarse en 10TB LUN con thin provisioning organizados en un grupo de discos de LVM. Ocuparía solo 2TB GB de espacio en el momento de la creación y solo reclamaría espacio adicional a medida que se asignan extensiones para acomodar el crecimiento de los datos. Este proceso es seguro siempre y cuando se supervise el espacio.

Recuperación tras fallos y cambio de ONTAP

Se requiere comprender las funciones de toma de control y conmutación de sitios de almacenamiento para garantizar que estas operaciones no interrumpan las operaciones de la base de datos de Oracle. Además, los argumentos utilizados por las operaciones de toma de control y conmutación de sitios pueden afectar a la integridad de los datos si se usan incorrectamente.

- En condiciones normales, las escrituras entrantes en una controladora determinada se reflejan de forma síncrona en su compañero. En un entorno NetApp MetroCluster, las escrituras también se reflejan en una controladora remota. No se reconoce en la aplicación host hasta que se almacena una escritura en medios no volátiles en todas las ubicaciones.
- El medio que almacena los datos de escritura se denomina memoria no volátil o NVMEM. También se conoce a veces como memoria de acceso aleatorio no volátil (NVRAM), y se puede considerar como una caché de escritura aunque funciona como un diario. En un funcionamiento normal, los datos de NVMEM no se leen; solo se utilizan para proteger los datos en caso de un fallo de software o hardware. Cuando se escriben datos en las unidades, los datos se transfieren desde la RAM del sistema, no desde NVMEM.
- Durante una operación de toma de control, un nodo de una pareja de alta disponibilidad toma el control de las operaciones de su compañero. Una conmutación de sitios es básicamente la misma, pero se aplica a las configuraciones de MetroCluster en las que un nodo remoto toma las funciones de un nodo local.

Durante las operaciones de mantenimiento rutinarias, una operación de toma de control o de conmutación de sitios debería ser transparente, excepto en una breve pausa potencial de las operaciones cuando cambian las rutas de red. Sin embargo, las redes pueden ser complicadas y es fácil cometer errores, por lo que NetApp recomienda encarecidamente probar exhaustivamente las operaciones de toma de control y conmutación antes de poner un sistema de almacenamiento en producción. Hacerlo es la única forma de asegurarse de que todas las rutas de red están configuradas correctamente. En un entorno SAN, compruebe cuidadosamente la salida del comando `sanlun lun show -p` para asegurarse de que todas las rutas primarias y secundarias esperadas estén disponibles.

Se debe tener cuidado al emitir una toma de control forzada o cambio. Al forzar un cambio en la configuración de almacenamiento con estas opciones, se ignorará el estado de la controladora propietaria de las unidades y el nodo alternativo tomará el control de las unidades de manera forzada. El forzado incorrecto de una toma de control puede provocar la pérdida de datos o la corrupción. Esto se debe a que una toma de control o una conmutación por error forzada pueden descartar el contenido de NVMEM. Una vez completada la toma de control o la conmutación por error, la pérdida de esos datos implica que los datos almacenados en las unidades pueden revertir a un estado ligeramente más antiguo desde el punto de vista de la base de datos.

En raras ocasiones se debería necesitar una toma de control forzada con un par de alta disponibilidad normal. En prácticamente todas las situaciones de fallo, un nodo se apaga e informa al partner para que se produzca una conmutación automática al respaldo. Hay algunos casos periféricos, como un fallo gradual en el que se pierde la interconexión entre nodos y después se pierde una controladora, en el que se requiere una toma de control forzada. En esta situación, el mirroring entre nodos se pierde antes del fallo de la controladora, lo que significa que la controladora superviviente ya no tendría una copia de las escrituras en curso. Entonces, se debe forzar la toma de control, lo que significa que potencialmente se pueden perder los datos.

La misma lógica se aplica a un switchover de MetroCluster. En condiciones normales, una conmutación es prácticamente transparente. Sin embargo, un desastre puede resultar en una pérdida de conectividad entre el

sitio sobreviviente y el sitio del desastre. Desde el punto de vista del sitio sobreviviente, el problema podría ser nada más que una interrupción en la conectividad entre sitios, y el sitio original podría aún estar procesando datos. Si un nodo no puede comprobar el estado de la controladora principal, solo es posible realizar una conmutación de sitios forzada.

NetApp recomienda tomar las siguientes precauciones:

- Tenga mucho cuidado de no forzar accidentalmente una toma de control o una conmutación de sitios. Normalmente, no se debe forzar, y forzar el cambio puede provocar la pérdida de datos.
- Si se requiere una toma de control forzada o una conmutación por error, asegúrese de que las aplicaciones estén cerradas, todos los sistemas de archivos estén desmontados y los grupos de volúmenes del gestor de volúmenes lógicos (LVM) se varyoffs. Los grupos de discos de ASM deben estar desmontados.
- En caso de una conmutación de MetroCluster forzada, elimine el nodo fallido de todos los recursos de almacenamiento que sobrevivan. Para obtener más información, consulte la Guía de gestión de MetroCluster y recuperación ante desastres para la versión relevante de ONTAP.

MetroCluster y varios agregados

MetroCluster es una tecnología de replicación síncrona que cambia al modo asíncrono en caso de interrupción de la conectividad. Esta es la solicitud más común de los clientes, porque la replicación síncrona garantizada implica que la interrupción de la conectividad del sitio provoca una parada completa de las operaciones de I/O de la base de datos, lo que impide que la base de datos funcione.

Con MetroCluster, los agregados se resincronizan rápidamente después de restaurar la conectividad. A diferencia de otras tecnologías de almacenamiento, MetroCluster nunca debería requerir un nuevo mirroring completo tras un fallo del sitio. Sólo se deben enviar los cambios delta.

En conjuntos de datos que abarcan agregados, existe el pequeño riesgo de que se requieran pasos adicionales de recuperación de datos en un escenario de desastre continuo. Específicamente, si (a) se interrumpe la conectividad entre sitios, (b) se restaura la conectividad, (c) los agregados alcanzan un estado en el que algunos están sincronizados y otros no, y luego (d) se pierde el sitio principal, el resultado es un sitio superviviente en el que los agregados no están sincronizados entre sí. Si esto sucede, algunas partes del conjunto de datos se sincronizan entre sí y no es posible activar aplicaciones, bases de datos o almacenes de datos sin recuperación. Si un conjunto de datos abarca agregados, NetApp recomienda aprovechar los backups basados en instantáneas con una de las muchas herramientas disponibles para verificar la capacidad de recuperación rápida en este escenario inusual.

Configuración de ONTAP en sistemas ASA r2

RAID

RAID se refiere al uso de redundancia basada en paridad para proteger los datos contra fallas de la unidad. ASA r2 utiliza las mismas tecnologías ONTAP RAID que los sistemas AFF y FAS , lo que garantiza una protección sólida contra fallas de múltiples discos.

ONTAP realiza la configuración RAID automáticamente para los sistemas ASA r2. Este es un componente central de la experiencia de administración de almacenamiento simplificada introducida con la personalidad ASA r2.

Los detalles clave sobre la configuración RAID automática en ASA r2 incluyen:

- Zonas de disponibilidad de almacenamiento (SAZ): en lugar de administrar manualmente los agregados tradicionales y los grupos RAID, ASA r2 utiliza zonas de disponibilidad de almacenamiento (SAZ). Se trata de grupos de discos compartidos y protegidos mediante RAID para un par de alta disponibilidad, donde ambos nodos tienen acceso completo al mismo almacenamiento.
- Ubicación automática: cuando se crea una unidad de almacenamiento (LUN o espacio de nombres NVMe), ONTAP crea automáticamente un volumen dentro de la SAZ y lo coloca para lograr un rendimiento y un equilibrio de capacidad óptimos.
- Sin administración manual de agregados: los comandos tradicionales de administración de agregados y grupos RAID no son compatibles con ASA r2. Esto elimina la necesidad de que los administradores planifiquen manualmente los tamaños de los grupos RAID, los discos de paridad o las asignaciones de nodos.
- Aprovisionamiento simplificado: el aprovisionamiento se maneja a través del Administrador del sistema o comandos CLI simplificados que se centran en las unidades de almacenamiento en lugar del diseño RAID físico subyacente.
- Reequilibrio de la carga de trabajo: a partir de las versiones 2025 (ONTAP 9.17.1), ONTAP reequilibra automáticamente las cargas de trabajo entre los nodos del par HA para garantizar que el rendimiento y la utilización del espacio permanezcan equilibrados sin intervención manual.

ASA r2 utiliza automáticamente las tecnologías RAID predeterminadas de ONTAP: RAID DP para la mayoría de las configuraciones y RAID-TEC para grupos de SSD muy grandes. Esto elimina la necesidad de selección RAID manual. Estos niveles RAID basados en paridad brindan una mejor confiabilidad y eficiencia de almacenamiento que la duplicación, que las prácticas recomendadas de Oracle más antiguas suelen recomendar, pero que no es relevante para ASA r2. ONTAP evita la penalización de escritura RAID tradicional a través de la integración de WAFL , lo que garantiza un rendimiento óptimo para las cargas de trabajo de Oracle, como el registro de rehacer y las escrituras aleatorias de archivos de datos. Combinado con la gestión RAID automatizada y las zonas de disponibilidad de almacenamiento, ASA r2 ofrece alta disponibilidad y protección de nivel empresarial para bases de datos Oracle.

Gestión de la capacidad

Gestionar una base de datos u otra aplicación empresarial con almacenamiento empresarial predecible, gestionable y de alto rendimiento requiere cierto espacio libre en las unidades para la gestión de datos y metadatos. La cantidad de espacio libre necesario depende del tipo de unidad utilizada y los procesos empresariales.

ASA r2 utiliza zonas de disponibilidad de almacenamiento (SAZ) en lugar de agregados, pero el principio sigue siendo el mismo: el espacio libre incluye cualquier capacidad física no consumida por datos reales, instantáneas o sobrecarga del sistema. También debe considerarse el aprovisionamiento fino: las asignaciones lógicas no reflejan el uso físico real.

Las recomendaciones de NetApp para los sistemas de almacenamiento ASA r2 utilizados para aplicaciones empresariales son las siguientes:

Grupos de SSD en sistemas ASA r2

* NetApp recomienda* mantener un mínimo del 10 % de espacio físico libre en entornos ASA r2. Esta guía se aplica a los grupos exclusivos de SSD utilizados por los sistemas ASA r2 e incluye todo el espacio no utilizado dentro de la SAZ y las unidades de almacenamiento. El espacio lógico no es importante; el enfoque está en el espacio físico libre real disponible para el almacenamiento de datos.

Si bien ASA r2 puede soportar un alto nivel de utilización sin degradación del rendimiento, operar cerca de la capacidad total aumenta el riesgo de agotamiento del espacio y de gastos administrativos al expandir el almacenamiento. Es posible que funcionar con más del 90 % de utilización no afecte el rendimiento, pero puede complicar la administración y retrasar el aprovisionamiento de unidades adicionales.

Los sistemas ASA r2 admiten unidades de almacenamiento de hasta 128 TB y tamaños de SAZ de hasta 2 PB por par HA, y ONTAP equilibra automáticamente la capacidad entre los nodos. Monitorear la utilización a nivel de clúster, SAZ y unidad de almacenamiento es esencial para garantizar suficiente espacio libre para instantáneas, cargas de trabajo con aprovisionamiento fino y crecimiento futuro. Si la capacidad se acerca a los umbrales críticos (~90 % de utilización), se deben agregar SSD adicionales en grupos (mínimo seis unidades) para mantener el rendimiento y la resiliencia.

Máquinas virtuales de almacenamiento

La gestión del almacenamiento de la base de datos de Oracle en los sistemas ASA r2 también está centralizada en una máquina virtual de almacenamiento (SVM), conocida como vserver en la CLI de ONTAP .

Una SVM es la unidad fundamental de aprovisionamiento y seguridad de almacenamiento en ONTAP, similar a una VM invitada en un servidor VMware ESX. Cuando ONTAP se instala por primera vez en ASA r2, no tiene capacidades de servicio de datos hasta que se crea una SVM. La SVM define la personalidad y los servicios de datos para el entorno SAN.

Los sistemas ASA r2 utilizan una personalidad ONTAP solo para SAN, que está optimizada para admitir protocolos de bloque (FC, iSCSI, NVMe/FC, NVMe/TCP) y elimina las funciones relacionadas con NAS. Esto simplifica la gestión y garantiza que todas las configuraciones de SVM estén optimizadas para las cargas de trabajo SAN. A diferencia de los sistemas AFF/ FAS , ASA r2 no expone opciones para servicios NAS como directorios de inicio o recursos compartidos NFS.

Cuando se crea un clúster, ASA r2 aprovisiona automáticamente un SVM de datos predeterminado llamado svm1 con protocolos SAN habilitados. Esta SVM está lista para operaciones de almacenamiento en bloque sin requerir configuración manual de servicios de protocolo. De forma predeterminada, los LIF de datos IP en esta SVM admiten los protocolos iSCSI y NVMe/TCP y utilizan la política de servicio de bloques de datos predeterminados, lo que simplifica la configuración inicial para cargas de trabajo SAN. Posteriormente, los administradores pueden crear SVM adicionales o personalizar configuraciones de LIF según los requisitos de rendimiento, seguridad o múltiples inquilinos.

Las interfaces lógicas (LIF) para los protocolos SAN deben diseñarse en función de los requisitos de rendimiento y disponibilidad. ASA r2 admite LIF iSCSI, FC y NVMe, pero tenga en cuenta que la conmutación por error automática de LIF iSCSI no está habilitada de manera predeterminada porque ASA r2 usa redes compartidas para hosts NVMe y SCSI. Para habilitar la conmutación por error automática, cree "LIF solo para iSCSI".

SVM

Al igual que con otras plataformas ONTAP , no existe una práctica recomendada oficial sobre la cantidad de SVM a crear; la decisión depende de los requisitos de administración y seguridad.

La mayoría de los clientes operan un único SVM principal para las operaciones diarias y crean SVM adicionales para necesidades especiales, como:

- Un SVM dedicado para una base de datos empresarial crítica administrada por un equipo de especialistas
- Una SVM para un grupo de desarrollo con control administrativo delegado
- Una SVM para datos confidenciales que requieren acceso administrativo restringido

En entornos de múltiples inquilinos, a cada inquilino se le puede asignar una SVM dedicada. El límite para la cantidad de SVM y LIF por clúster, par de HA y nodo depende del protocolo utilizado, el modelo de nodo y la versión de ONTAP. Consultar el "Hardware Universe de NetApp" para estos límites.

ASA r2 admite hasta 256 SVM por clúster y por par de alta disponibilidad a partir de ONTAP 9.18.1 (anteriormente 32 en versiones anteriores).

Gestión del rendimiento con ONTAP QoS en sistemas ASA r2

Para gestionar de forma segura y eficiente varias bases de datos Oracle en ASA r2 se requiere una estrategia de calidad de servicio eficaz. Esto es especialmente importante porque los sistemas ASA r2 son plataformas SAN totalmente flash diseñadas para un rendimiento extremadamente alto y una consolidación de carga de trabajo.

Una cantidad relativamente pequeña de SSD puede saturar incluso los controladores más potentes, por lo que los controles de QoS son esenciales para garantizar un rendimiento predecible en múltiples cargas de trabajo. Como referencia, los sistemas ASA r2 como ASA A1K o A90 pueden ofrecer de cientos de miles a más de un millón de IOPS con una latencia de submilisegundos. Muy pocas cargas de trabajo individuales consumirían este nivel de rendimiento, por lo que la utilización completa generalmente implica alojar múltiples bases de datos o aplicaciones. Para hacer esto de forma segura se requieren políticas de calidad de servicio para evitar la contención de recursos.

ONTAP QoS en ASA r2 funciona de la misma manera que en los sistemas AFF/ FAS , con dos tipos principales de controles: IOPS y ancho de banda. Los controles de QoS se pueden aplicar a SVM y LUN.

Calidad de servicio IOPS

La calidad de servicio basada en IOPS limita la cantidad total de IOPS para un recurso determinado. En ASA r2, las políticas de QoS se pueden aplicar a nivel de SVM y a objetos de almacenamiento individuales, como LUN. Cuando una carga de trabajo alcanza su límite de IOPS, se crean colas de solicitudes de E/S adicionales para obtener tokens, lo que genera latencia. Este es el comportamiento esperado y evita que una sola carga de trabajo monopolice los recursos del sistema.

Tenga cuidado al aplicar controles de QoS a datos de registros de transacciones/rehacer de bases de datos. Estas cargas de trabajo son intermitentes, y un límite de QoS que parece razonable para una actividad promedio puede ser demasiado bajo para picos de actividad, lo que causa graves problemas de rendimiento. En general, el registro de rehacer y archivar no debería estar limitado por la calidad de servicio.

Calidad del ancho de banda

La calidad de servicio basada en ancho de banda limita el rendimiento en Mbps. Esto es útil cuando las cargas de trabajo realizan lecturas o escrituras de bloques grandes, como escaneos de tablas completas u operaciones de respaldo, que consumen un ancho de banda significativo pero relativamente pocas IOPS. La

combinación de límites de IOPS y ancho de banda puede proporcionar un control más granular.

Calidad de servicio mínima/garantizada

Las políticas de QoS mínimas reservan el rendimiento para cargas de trabajo críticas. Por ejemplo, en un entorno mixto con bases de datos de producción y desarrollo, aplique la máxima QoS a las cargas de trabajo de desarrollo y la mínima QoS a las cargas de trabajo de producción para garantizar un rendimiento predecible.

Calidad de servicio adaptativa

La calidad de servicio adaptativa ajusta los límites en función del tamaño del objeto de almacenamiento. Si bien rara vez se utiliza para bases de datos (porque el tamaño no se correlaciona con las necesidades de rendimiento), puede ser útil para cargas de trabajo de virtualización donde los requisitos de rendimiento escalan con la capacidad.

Eficiencia

Las funciones de eficiencia espacial de ONTAP son totalmente compatibles y optimizadas para los sistemas ASA r2. En casi todos los casos, el mejor enfoque es dejar los valores predeterminados con todas las funciones de eficiencia habilitadas.

Los sistemas ASA r2 son plataformas SAN totalmente flash, por lo que las tecnologías de eficiencia como la compresión, la compactación y la deduplicación son fundamentales para maximizar la capacidad utilizable y reducir los costos.

Compresión

La compresión reduce los requisitos de espacio al codificar patrones en los datos. Con los sistemas ASA r2 basados en SSD, la compresión ofrece ahorros significativos porque flash elimina la necesidad de aprovisionamiento excesivo para mejorar el rendimiento. La compresión adaptativa de ONTAP está habilitada de forma predeterminada y se ha probado exhaustivamente con cargas de trabajo empresariales, incluidas bases de datos Oracle, sin un impacto medible en el rendimiento, incluso en entornos donde la latencia se mide en microsegundos. En algunos casos, el rendimiento mejora porque los datos comprimidos ocupan menos espacio en caché.

> La eficiencia de almacenamiento sensible a la temperatura (TSSE) no se aplica en los sistemas ASA r2. En los sistemas ASA r2, la compresión no se basa en datos activos (a los que se accede con frecuencia) ni en datos inactivos (a los que se accede con poca frecuencia). La compresión comienza sin esperar a que los datos se enfríen.

Compresión adaptativa

La compresión adaptativa utiliza un tamaño de bloque de 8 KB de forma predeterminada, que coincide con el tamaño de bloque comúnmente utilizado por las bases de datos relacionales. Los tamaños de bloque más grandes (16 KB o 32 KB) pueden mejorar la eficiencia de los datos secuenciales, como registros de transacciones o copias de seguridad, pero se deben utilizar con precaución en bases de datos activas para evitar sobrecarga durante las sobrescrituras.

> El tamaño del bloque se puede aumentar hasta 32 KB para archivos inactivos, como registros o copias de seguridad. Consulte la guía de NetApp antes de cambiar los valores predeterminados.

No utilice compresión de 32 KB con deduplicación para copias de seguridad en streaming. Utilice una compresión de 8 KB para mantener la eficiencia de la deduplicación.

Alineación de la compresión

La alineación de la compresión es importante para las sobrescrituras aleatorias. Asegúrese de que el tipo de LUN, el desplazamiento de partición (múltiplo de 8 KB) y el tamaño del bloque del sistema de archivos sean correctos y estén alineados con el tamaño del bloque de la base de datos. Los datos secuenciales, como copias de seguridad o registros, no requieren consideraciones de alineación.

Compactación de datos

La compactación complementa la compresión al permitir que varios bloques comprimidos compartan el mismo bloque físico. Por ejemplo, si un bloque de 8 KB se comprime a 1 KB, la compactación garantiza que el espacio restante no se desperdicie. Esta función está en línea y no introduce penalizaciones en el rendimiento.

Deduplicación

La deduplicación elimina bloques duplicados en conjuntos de datos. Si bien las bases de datos Oracle generalmente producen ahorros mínimos en la deduplicación debido a los encabezados y finales de bloques únicos, la deduplicación ONTAP aún puede recuperar espacio de bloques en cero y patrones repetidos.

Eficiencia y thin provisioning

Los sistemas ASA r2 utilizan aprovisionamiento fino de forma predeterminada. Las características de eficiencia complementan el aprovisionamiento fino para maximizar la capacidad utilizable.

Las unidades de almacenamiento siempre tienen un aprovisionamiento fino en los sistemas de almacenamiento ASA r2. No se admite el aprovisionamiento grueso.

Tecnología QuickAssist (QAT)

En las plataformas NetApp ASA r2, la tecnología Intel QuickAssist (QAT) proporciona una eficiencia acelerada por hardware que difiere significativamente de la eficiencia de almacenamiento sensible a la temperatura (TSSE) basada en software sin QAT.

QAT con aceleración de hardware:

- Descarga tareas de compresión y cifrado de los núcleos de la CPU.
- Permite una eficiencia inmediata y en línea tanto para datos activos (de acceso frecuente) como inactivos (de acceso poco frecuente).
- Reduce significativamente la sobrecarga de la CPU.
- Ofrece mayor rendimiento y menor latencia.
- Mejora la escalabilidad para operaciones sensibles al rendimiento, como el cifrado TLS y VPN.

TSSE sin QAT:

- Se basa en procesos controlados por CPU para operaciones de eficiencia.
- Aplica eficiencia sólo a datos fríos después de un retraso.

- Consume más recursos de la CPU.
- Limita el rendimiento general en comparación con los sistemas acelerados por QAT.

Por lo tanto, los sistemas ASA r2 modernos ofrecen una eficiencia más rápida y acelerada por hardware y una mejor utilización del sistema que las plataformas anteriores que solo utilizaban TSSE.

Mejores prácticas de eficiencia para ASA r2

NetApp recomienda lo siguiente:

Valores predeterminados de ASA r2

Las unidades de almacenamiento creadas en ONTAP que se ejecutan en sistemas ASA r2 tienen aprovisionamiento fino con todas las funciones de eficiencia en línea habilitadas de manera predeterminada, incluidas la compresión, la compactación y la deduplicación. Aunque las bases de datos Oracle generalmente no se benefician significativamente de la deduplicación y pueden incluir datos no comprimibles, estos valores predeterminados son apropiados para casi todas las cargas de trabajo. ONTAP está diseñado para procesar eficientemente todo tipo de datos y patrones de E/S, independientemente de que resulten en ahorros o no. Los valores predeterminados solo se deben cambiar si se comprenden plenamente los motivos y existe un beneficio claro en desviarse.

Recomendaciones generales

- Deshabilitar la compresión de datos cifrados o comprimidos por la aplicación: si los archivos ya están comprimidos en el nivel de la aplicación o cifrados, deshabilite la compresión para optimizar el rendimiento y permitir un funcionamiento más eficiente en otras unidades de almacenamiento.
- Evite combinar bloques de compresión grandes con deduplicación: no utilice compresión de 32 KB y deduplicación para las copias de seguridad de bases de datos. Para realizar copias de seguridad en streaming, utilice una compresión de 8 KB para mantener la eficiencia de la deduplicación.
- Supervise los ahorros de eficiencia: utilice las herramientas ONTAP (System Manager, Active IQ) para realizar un seguimiento de los ahorros de espacio reales y ajustar las políticas si es necesario.

Aprovisionamiento ligero

El aprovisionamiento fino para una base de datos Oracle en ASA r2 requiere una planificación cuidadosa porque implica configurar más espacio lógico del que está físicamente disponible. Cuando se implementa correctamente, el aprovisionamiento fino genera importantes ahorros de costos y una mejor capacidad de administración.

El aprovisionamiento fino es parte integral de ASA r2 y está estrechamente relacionado con las tecnologías de eficiencia ONTAP porque ambos permiten almacenar más datos lógicos que la capacidad física del sistema. Los sistemas ASA r2 son solo SAN y el aprovisionamiento fino se aplica a unidades de almacenamiento y LUN dentro de zonas de disponibilidad de almacenamiento (SAZ).

> Las unidades de almacenamiento ASA r2 tienen aprovisionamiento fino de manera predeterminada.

Casi cualquier uso de instantáneas implica aprovisionamiento fino. Por ejemplo, una base de datos típica de 10 TiB con 30 días de instantáneas podría aparecer como 310 TiB de datos lógicos, pero solo se consumen entre 12 TiB y 15 TiB de espacio físico porque las instantáneas solo almacenan bloques modificados.

De manera similar, la clonación es otra forma de aprovisionamiento fino. Un entorno de desarrollo con 40

clones de una base de datos de 80 TiB requeriría 3,2 PiB si estuviera completamente escrito, pero en la práctica consume mucho menos porque solo se almacenan los cambios.

Gestión del espacio

Se debe tener cierto cuidado con el aprovisionamiento fino en un entorno de aplicaciones porque las tasas de cambio de datos pueden aumentar inesperadamente. Por ejemplo, el consumo de espacio debido a las instantáneas puede crecer rápidamente si se reindexan las tablas de bases de datos o se aplican parches a gran escala a los invitados de VMware. Una copia de seguridad fuera de lugar puede escribir una gran cantidad de datos en muy poco tiempo. Por último, puede resultar difícil recuperar algunas aplicaciones si un LUN se queda sin espacio libre inesperadamente.

En ASA r2, estos riesgos se mitigan mediante **aprovisionamiento fino**, **monitoreo proactivo** y **políticas de cambio de tamaño de LUN**, en lugar de funciones de ONTAP como crecimiento automático de volúmenes o eliminación automática de instantáneas. Los administradores deben:

- Habilitar el aprovisionamiento fino en LUN (`space-reserve disabled`) - esta es la configuración predeterminada en ASA r2
- Supervise la capacidad mediante alertas de System Manager o automatización basada en API
- Utilice el cambio de tamaño de LUN programado o con script para adaptarse al crecimiento
- Configurar la reserva de instantáneas y la eliminación automática de instantáneas a través del Administrador del sistema (GUI)

> La planificación cuidadosa de los umbrales de espacio y los scripts de automatización es esencial porque ASA r2 no admite el crecimiento automático del volumen ni la eliminación de instantáneas impulsada por CLI.

ASA r2 no utiliza configuraciones de reserva fraccionaria porque es una arquitectura solo SAN que abstrae las opciones de volumen basadas en WAFL. En cambio, la eficiencia del espacio y la protección contra sobrescritura se gestionan a nivel de LUN. Por ejemplo, si tiene un LUN de 250 GiB aprovisionado desde una unidad de almacenamiento, las instantáneas consumen espacio en función de los cambios de bloque reales en lugar de reservar una cantidad igual de espacio por adelantado. Esto elimina la necesidad de grandes reservas estáticas, que eran comunes en los entornos ONTAP tradicionales que utilizaban reserva fraccionaria.

> Si se requiere protección de sobrescritura garantizada y la supervisión no es posible, los administradores deben aprovisionar capacidad suficiente en la unidad de almacenamiento y configurar la reserva de instantáneas de forma adecuada. Sin embargo, el diseño de ASA r2 hace que la reserva fraccionaria sea innecesaria para la mayoría de las cargas de trabajo.

Compresión y deduplicación

La compresión y la deduplicación en ASA r2 son tecnologías de eficiencia espacial, no mecanismos tradicionales de aprovisionamiento fino. Estas características reducen el espacio de almacenamiento físico al eliminar datos redundantes y comprimir bloques, lo que permite almacenar más datos lógicos de los que la capacidad bruta permitiría de otro modo.

Por ejemplo, un conjunto de datos de 50 TiB podría comprimirse a 30 TiB, ahorrando 20 TiB de espacio físico. Desde la perspectiva de la aplicación, todavía hay 50 TiB de datos, aunque ocupa solo 30 TiB en el disco.

La compresibilidad de un conjunto de datos puede cambiar con el tiempo, lo que puede aumentar el consumo de espacio físico. Por lo tanto, la compresión y la deduplicación deben gestionarse de forma proactiva mediante la supervisión y la planificación de la capacidad.

Espacio libre y asignación de espacio LVM

El aprovisionamiento fino en entornos ASA r2 puede perder eficiencia con el tiempo si los bloques eliminados no se recuperan. A menos que se libere espacio usando TRIM/UNMAP o se sobrescriba con ceros (a través de ASMRU, Utilidad de recuperación y administración automática de espacio), los datos eliminados continúan consumiendo capacidad física. En muchos entornos de bases de datos Oracle, el aprovisionamiento fino ofrece beneficios limitados porque los archivos de datos generalmente se asignan previamente a su tamaño completo durante la creación.

Una planificación cuidadosa de la configuración de LVM puede mejorar la eficiencia y minimizar la necesidad de aprovisionamiento de almacenamiento y cambio de tamaño de LUN. Cuando se utiliza un LVM como Veritas VxVM u Oracle ASM, los LUN subyacentes se dividen en extensiones que solo se utilizan cuando es necesario. Por ejemplo, si un conjunto de datos comienza con un tamaño de 2 TiB pero podría crecer hasta 10 TiB con el tiempo, este conjunto de datos podría ubicarse en 10 TiB de LUN con aprovisionamiento fino organizados en un grupo de discos LVM. Ocuparía solo 2 TiB de espacio en el momento de la creación y solo reclamaría espacio adicional a medida que se asignen extensiones para acomodar el crecimiento de los datos. Este proceso es seguro siempre que se controle el espacio.

Conmutación por error de ONTAP

Es necesario comprender las funciones de adquisición de almacenamiento para garantizar que las operaciones de la base de datos Oracle no se interrumpan durante estas operaciones. Además, los argumentos utilizados en las operaciones de adquisición pueden afectar la integridad de los datos si se utilizan incorrectamente.

En condiciones normales, las escrituras entrantes a un controlador determinado se reflejan de manera sincrónica en su socio de alta disponibilidad. En un entorno ASA r2 con SnapMirror Active Sync (SM-as), las escrituras también se reflejan en un controlador remoto en el sitio secundario. Hasta que una escritura no se almacena en un medio no volátil en todas las ubicaciones, no se reconoce en la aplicación host.

El medio que almacena los datos escritos se llama memoria no volátil (NVMEM). A veces se la denomina memoria de acceso aleatorio no volátil (NVRAM) y puede considerarse como un diario de escritura en lugar de un caché. Durante el funcionamiento normal, los datos de NVMEM no se leen; solo se utilizan para proteger los datos en caso de una falla de software o hardware. Cuando se escriben datos en las unidades, los datos se transfieren desde la RAM del sistema, no desde NVMEM.

Durante una operación de adquisición, un nodo de un par de HA se hace cargo de las operaciones de su socio. En ASA r2, el cambio no es aplicable porque no se admite MetroCluster ; en su lugar, SnapMirror Active Sync proporciona redundancia a nivel de sitio. Las operaciones de toma de control de almacenamiento durante el mantenimiento de rutina deben ser transparentes, salvo una breve pausa en las operaciones a medida que cambian las rutas de la red. Las redes pueden ser complejas y es fácil cometer errores, por lo que NetApp recomienda encarecidamente probar exhaustivamente las operaciones de adquisición antes de poner un sistema de almacenamiento en producción. Hacerlo es la única forma de garantizar que todas las rutas de red estén configuradas correctamente. En un entorno SAN, verifique el estado de la ruta mediante el comando `sanlun lun show -p` o las herramientas de múltiples rutas nativas del sistema operativo para garantizar que todas las rutas esperadas estén disponibles. Los sistemas ASA r2 proporcionan todas las rutas optimizadas activas para LUN, y los clientes que usan espacios de nombres NVMe deben confiar en herramientas nativas del sistema operativo, ya que sanlun no cubre las rutas NVMe.

Se debe tener cuidado al emitir una adquisición forzosa. Forzar un cambio en la configuración de almacenamiento significa que se ignora el estado del controlador que posee las unidades y el nodo alternativo toma el control de las unidades de manera forzosa. La forzamiento incorrecto de una toma de control puede provocar pérdida o corrupción de datos porque una toma de control forzada puede descartar el contenido de NVMEM. Una vez completada la adquisición, la pérdida de esos datos significa que los datos almacenados en las unidades podrían volver a un estado ligeramente más antiguo desde el punto de vista de la base de datos.

En raras ocasiones debería ser necesaria una adquisición forzada con un par HA normal. En casi todos los escenarios de falla, un nodo se apaga e informa al socio para que se realice una conmutación por error automática. Hay algunos casos extremos, como una falla continua en la que se pierde la interconexión entre los nodos y luego falla un controlador, en los que se requiere una toma de control forzada. En tal situación, la duplicación entre nodos se pierde antes de la falla del controlador, lo que significa que el controlador sobreviviente ya no tiene una copia de las escrituras en progreso. Luego es necesario forzar la toma de control, lo que potencialmente implica una pérdida de datos.

NetApp recomienda tomar las siguientes precauciones:

- Tenga mucho cuidado de no forzar una adquisición accidentalmente. Normalmente no debería ser necesario forzar el cambio, ya que forzarlo puede provocar la pérdida de datos.
- Si se requiere una toma de control forzada, asegúrese de que las aplicaciones estén cerradas, todos los sistemas de archivos estén desmontados y los grupos de volúmenes del administrador de volúmenes lógicos (LVM) estén desactivados. Los grupos de discos ASM deben desmontarse.
- En el caso de una falla a nivel de sitio cuando se utilizan SM-as, se iniciará la conmutación por error automática no planificada asistida por ONTAP Mediator en el clúster sobreviviente, lo que provocará una breve pausa de E/S y luego las transiciones de la base de datos continuarán desde el clúster sobreviviente. Para obtener más información, consulte la "Sincronización activa de SnapMirror en sistemas ASA r2" para conocer los pasos de configuración detallados.

Configuración de bases de datos con sistemas AFF/ FAS

Tamaños de bloque

ONTAP utiliza internamente un tamaño de bloque variable, lo que significa que las bases de datos Oracle se pueden configurar con el tamaño de bloque deseado. Sin embargo, los tamaños de bloque del sistema de archivos pueden afectar al rendimiento y, en algunos casos, un tamaño de bloque de redo más grande puede mejorar el rendimiento.

Tamaños de bloque de archivos de datos

Algunos sistemas operativos ofrecen diferentes tamaños de bloque del sistema de archivos. En el caso de los sistemas de archivos que admiten archivos de datos de Oracle, el tamaño de bloque debe ser 8KB cuando se utiliza la compresión. Cuando no se necesita compresión, se puede utilizar un tamaño de bloque de 8KB KB o 4KB KB.

Si se coloca un archivo de datos en un sistema de archivos con un bloque de 512 bytes, es posible que los archivos estén mal alineados. El LUN y el sistema de archivos podrían alinearse correctamente de acuerdo con las recomendaciones de NetApp, pero la I/O de archivo estaría mal alineada. Tal desalineación podría causar graves problemas de rendimiento.

Los sistemas de archivos compatibles con redo logs deben utilizar un tamaño de bloque que sea múltiplo del tamaño del bloque de redo. Esto generalmente requiere que tanto el sistema de archivos redo log como el propio redo log utilicen un tamaño de bloque de 512 bytes.

Rehacer tamaños de bloques

Con tasas de redo muy elevadas, es posible que los bloques de 4KB KB rindan mejor porque las tasas de rehacer elevadas permiten realizar I/O en operaciones cada vez más eficientes. Si las tasas de redo son mayores que 50Mbps, considere la posibilidad de probar un tamaño de bloque de 4KB KB.

Se han identificado algunos problemas de los clientes con bases de datos que utilizan redo logs con un tamaño de bloque de 512 bytes en un sistema de archivos con un tamaño de bloque de 4KB KB y muchas transacciones muy pequeñas. La sobrecarga involucrada en la aplicación de varios cambios de 512 bytes a un único bloque del sistema de archivos de 4KB se tradujo en problemas de rendimiento que se resolvieron mediante el cambio del sistema de archivos para que utilizara un tamaño de bloque de 512 bytes.

NetApp recomienda que no cambie el tamaño del bloque de redo a menos que se lo indique un servicio de atención al cliente relevante o una organización de servicios profesionales o que el cambio se base en la documentación oficial del producto.

db_file_multibloque_read_count

La `db_file_multiblock_read_count` El parámetro controla el Núm. Máximo de bloques de bases de datos Oracle que Oracle lee como una sola operación durante la E/S secuencial

Sin embargo, este parámetro no afecta a la cantidad de bloques que Oracle lee durante cualquier operación de lectura ni afecta a las operaciones de lectura aleatorias Solo se ve afectado el tamaño de bloque de I/O secuencial.

Oracle recomienda que el usuario deje este parámetro sin definir. Al hacerlo, el software de la base de datos puede definir automáticamente el valor óptimo. Por lo general, este parámetro se establece en un valor que proporciona un tamaño de I/O de 1MB. Por ejemplo, una lectura de 1MB de bloques de 8KB requeriría la lectura de 128 bloques y el valor predeterminado de este parámetro sería, por lo tanto, de 128.

La mayoría de los problemas de rendimiento de la base de datos observados por NetApp en los sitios de los clientes implican una configuración incorrecta para este parámetro. Hay motivos válidos para cambiar este valor con las versiones 8 y 9 de Oracle. Como resultado, el parámetro puede estar presente sin saberlo en `init.ora` Archivos porque la base de datos se actualizó in situ a Oracle 10 y versiones posteriores. Una configuración heredada de 8 o 16, en comparación con el valor predeterminado de 128, daña significativamente el rendimiento de I/O secuencial.

NetApp recomienda configurar el `db_file_multiblock_read_count` el parámetro no debe estar presente en el `init.ora` archivo. NetApp nunca se ha encontrado con una situación en la que cambiar este parámetro mejoró el rendimiento, pero hay muchos casos en los que causó daños claros en el rendimiento de I/O secuencial.

filesystemio_options

Parámetro de inicialización de Oracle `filesystemio_options` Controla el uso de la E/S asíncrona y directa

Contrariamente a la creencia común, las E/S asincrónicas y directas no son mutuamente excluyentes. NetApp ha observado que este parámetro suele estar mal configurado en los entornos del cliente, y esta mala configuración es el responsable directo de muchos problemas de rendimiento.

La E/S asíncrona significa que las operaciones de I/O de Oracle se pueden paralelizar. Antes de la disponibilidad de E/S asíncrona en varios sistemas operativos, los usuarios configuraron numerosos procesos de escritura de base de datos y cambiaron la configuración del proceso del servidor. Con la E/S asíncrona, el propio sistema operativo realiza E/S en nombre del software de base de datos de forma paralela y altamente eficiente. Este proceso no pone los datos en riesgo y las operaciones críticas, como el redo registro de Oracle, se siguen realizando de forma síncrona.

La E/S directa omite la caché de buffers del SO. Las E/S en un sistema UNIX normalmente fluyen a través de la caché de buffers del sistema operativo. Esto es útil para aplicaciones que no mantienen una caché interna, pero Oracle tiene su propia caché de buffers en SGA. En casi todos los casos, es mejor habilitar la E/S directa y asignar la RAM del servidor a la SGA en lugar de confiar en la caché de buffers del SO. Oracle SGA utiliza la memoria de forma más eficaz. Además, cuando la I/O fluye por el búfer del SO, se somete a un procesamiento adicional, lo que aumenta las latencias. El aumento de las latencias es especialmente notable en operaciones pesadas de I/O de escritura cuando un requisito crucial es la baja latencia.

Las opciones para `filesystemio_options` son:

- **Async.** Oracle envía solicitudes de E/S al sistema operativo para su procesamiento. Este proceso permite a Oracle realizar otro trabajo en lugar de esperar la finalización de E/S y, por lo tanto, aumenta la paralelización de E/S.
- **Directio.** Oracle realiza E/S directamente contra archivos físicos en lugar de enrutar E/S a través de la caché del SO host.
- **None.** Oracle utiliza E/S síncronas y en buffer En esta configuración, la elección entre los procesos de servidor compartido y dedicado y el número de dbwriters son más importantes.
- **Setall.** Oracle utiliza E/S tanto asíncrona como directa En casi todos los casos, el uso de `setall` es óptimo.

La `filesystemio_options` El parámetro no tiene ningún efecto en los entornos DNFS y ASM. El uso de DNFS o ASM da como resultado el uso de E/S tanto asíncrona como directa

Algunos clientes se han encontrado con problemas de E/S asíncronos en el pasado, especialmente con versiones anteriores de Red Hat Enterprise Linux 4 (RHEL4). Algunos consejos anticuados en Internet todavía sugieren evitar la IO asíncrona debido a la información obsoleta. La E/S asíncrona es estable en todos los sistemas operativos actuales. No hay motivo para desactivarlo, sin un error conocido en el sistema operativo.

Si una base de datos ha estado utilizando E/S en búfer, un cambio a E/S directa también puede justificar un cambio en el tamaño de SGA. Al desactivar las E/S en buffer, se elimina la ventaja de rendimiento que proporciona la caché del SO del host para la base de datos. Al volver a agregar RAM a SGA se soluciona este problema. El resultado neto debe ser una mejora en el rendimiento de E/S.

Aunque casi siempre es mejor utilizar RAM para Oracle SGA que para el almacenamiento en caché de buffers del sistema operativo, puede ser imposible determinar el mejor valor. Por ejemplo, puede ser preferible utilizar E/S en buffer con tamaños SGA muy pequeños en un servidor de bases de datos con muchas instancias de Oracle activas de forma intermitente. Esta disposición permite el uso flexible de la RAM libre restante en el SO por todas las instancias de base de datos en ejecución. Se trata de una situación muy inusual, pero se ha observado en algunos sitios de clientes.

NetApp recomienda ajuste `filesystemio_options` para `setall`, Pero tenga en cuenta que, en algunas circunstancias, la pérdida de la caché de buffers del host puede requerir un aumento en Oracle SGA.

Tiempo de espera de RAC

Oracle RAC es un producto de clusterware con varios tipos de procesos internos de latido que controlan el estado del cluster.

La información de la "recuento de errores" La sección incluye información crítica para entornos Oracle RAC que utilizan almacenamiento en red y, en muchos casos, la configuración predeterminada de Oracle RAC deberá cambiarse para garantizar que el cluster RAC sobrevive los cambios de ruta de red y las operaciones de failover/switchover de almacenamiento.

tiempo de espera del disco

El parámetro de RAC relacionado con el almacenamiento primario es `disktimeout`. Este parámetro controla el umbral en el que debe completarse la E/S del archivo de quorum. Si la `disktimeout` Se supera el parámetro y el nodo RAC se expulsa del clúster. El valor predeterminado de este parámetro es 200. Este valor debería ser suficiente para los procedimientos estándar de toma de control y devolución del almacenamiento.

NetApp recomienda probar exhaustivamente las configuraciones de RAC antes de colocarlas en producción, ya que existen muchos factores que afectan a una toma de control o al retorno primario. Además del tiempo necesario para que se complete la conmutación por error del almacenamiento, también se requiere más tiempo para que se propaguen los cambios del protocolo de control de agregación de enlaces (LACP). Además, el software multivía SAN debe detectar un tiempo de espera de I/O y volver a intentarlo en una ruta alternativa. Si una base de datos está extremadamente activa, se debe poner en cola una gran cantidad de E/S y volver a intentarlo antes de procesar la E/S del disco de quorum.

Si no se puede realizar una toma de control o una devolución del almacenamiento real, el efecto se puede simular mediante pruebas de extracción de cables en el servidor de bases de datos.

NetApp recomienda lo siguiente:

- Dejando el `disktimeout` parámetro con el valor predeterminado de 200.
- Pruebe siempre a fondo una configuración de RAC.

recuento de errores

La `misscount` Normalmente, el parámetro sólo afecta al latido de red entre los nodos de RAC. El valor predeterminado es 30 segundos. Si los binarios de grid se encuentran en una cabina de almacenamiento o si la unidad de arranque del sistema operativo no es local, este parámetro puede volverse importante. Esto incluye hosts con unidades de arranque ubicadas en una SAN FC, sistemas operativos arrancados NFS y unidades de arranque ubicadas en almacenes de datos de virtualización, como un archivo VMDK.

Si el acceso a una unidad de arranque se interrumpe por una toma de control o una restauración del almacenamiento, es posible que la ubicación binaria del grid o todo el sistema operativo se bloquee temporalmente. El tiempo necesario para que ONTAP complete la operación de almacenamiento y que el sistema operativo cambie de rutas y reanude las I/O puede superar el `misscount` umbral. Como resultado, un nodo se expulsa inmediatamente después de restaurar la conectividad con el LUN de arranque o los binarios de grid. En la mayoría de los casos, la expulsión y el reinicio posterior se producen sin mensajes de

registro que indiquen el motivo del reinicio. No todas las configuraciones se ven afectadas, por lo que debe realizar pruebas de cualquier host basado en almacenes de datos, arranque en NFS o arranque en SAN en un entorno RAC para que RAC se mantenga estable si se interrumpe la comunicación con la unidad de arranque.

En el caso de unidades de arranque no locales o de un sistema de archivos no local `grid` binarios, el `misscount` será necesario cambiar para que coincida `disktimeout`. Si se cambia este parámetro, realice otras pruebas para identificar también cualquier efecto sobre el comportamiento de RAC, como el tiempo de conmutación por error del nodo.

NetApp recomienda lo siguiente:

- Abandone el `misscount` parámetro con el valor por defecto de 30 a menos que se aplique una de las siguientes condiciones:
  - `grid` Los binarios se encuentran en una unidad conectada a la red, como las unidades basadas en almacén de datos, NFS, iSCSI y FC.
  - El sistema operativo se inicia mediante SAN.
- En tales casos, evalúe el efecto de las interrupciones de la red que afectan el acceso al sistema operativo o. `GRID_HOME` sistemas de ficheros: En algunos casos, estas interrupciones provocan que los daemons de Oracle RAC se atasquen, lo que puede provocar un `misscount`-basado en tiempo de espera y desalojo. El tiempo de espera predeterminado es 27 segundos, que es el valor de `misscount` menos `reboottime`. En tales casos, aumentar `misscount` a 200 para coincidir `disktimeout`.

Configuración de bases de datos con sistemas ASA r2

Tamaños de bloque

ONTAP utiliza internamente un tamaño de bloque variable, lo que significa que las bases de datos Oracle pueden configurarse con cualquier tamaño de bloque deseado. Sin embargo, los tamaños de los bloques del sistema de archivos pueden afectar el rendimiento y, en algunos casos, un tamaño de bloque de rehacer más grande puede mejorar el rendimiento.

ASA r2 no introduce ningún cambio en las recomendaciones de tamaño de bloque de Oracle en comparación con los sistemas AFF/ FAS . El comportamiento de ONTAP sigue siendo consistente en todas las plataformas.

Tamaños de bloque de archivos de datos

Algunos sistemas operativos ofrecen diferentes tamaños de bloque del sistema de archivos. En el caso de los sistemas de archivos que admiten archivos de datos de Oracle, el tamaño de bloque debe ser 8KB cuando se utiliza la compresión. Cuando no se necesita compresión, se puede utilizar un tamaño de bloque de 8KB KB o 4KB KB.

Si se coloca un archivo de datos en un sistema de archivos con un bloque de 512 bytes, es posible que los archivos estén mal alineados. El LUN y el sistema de archivos podrían alinearse correctamente de acuerdo con las recomendaciones de NetApp, pero la I/O de archivo estaría mal alineada. Tal desalineación podría causar graves problemas de rendimiento.

Rehacer tamaños de bloques

Los sistemas de archivos compatibles con redo logs deben utilizar un tamaño de bloque que sea múltiplo del tamaño del bloque de redo. Esto generalmente requiere que tanto el sistema de archivos redo log como el propio redo log utilicen un tamaño de bloque de 512 bytes.

Con tasas de redo muy elevadas, es posible que los bloques de 4KB KB rindan mejor porque las tasas de rehacer elevadas permiten realizar I/O en operaciones cada vez más eficientes. Si las tasas de redo son mayores que 50Mbps, considere la posibilidad de probar un tamaño de bloque de 4KB KB.

Se han identificado algunos problemas de los clientes con bases de datos que utilizan redo logs con un tamaño de bloque de 512 bytes en un sistema de archivos con un tamaño de bloque de 4KB KB y muchas transacciones muy pequeñas. La sobrecarga involucrada en la aplicación de varios cambios de 512 bytes a un único bloque del sistema de archivos de 4KB se tradujo en problemas de rendimiento que se resolvieron mediante el cambio del sistema de archivos para que utilizara un tamaño de bloque de 512 bytes.

NetApp recomienda que no cambie el tamaño del bloque de redo a menos que se lo indique un servicio de atención al cliente relevante o una organización de servicios profesionales o que el cambio se base en la documentación oficial del producto.

db_file_multibloque_read_count

La `db_file_multiblock_read_count` El parámetro controla el Núm. Máximo de bloques de bases de datos Oracle que Oracle lee como una sola operación durante la E/S secuencial

No hay cambios en las recomendaciones en comparación con los sistemas AFF/ FAS . El comportamiento de ONTAP y las mejores prácticas de Oracle siguen siendo idénticas en las plataformas ASA r2, AFF y FAS .

Sin embargo, este parámetro no afecta a la cantidad de bloques que Oracle lee durante cualquier operación de lectura ni afecta a las operaciones de lectura aleatorias Solo se ve afectado el tamaño de bloque de I/O secuencial.

Oracle recomienda que el usuario deje este parámetro sin definir. Al hacerlo, el software de la base de datos puede definir automáticamente el valor óptimo. Por lo general, este parámetro se establece en un valor que proporciona un tamaño de I/O de 1MB. Por ejemplo, una lectura de 1MB de bloques de 8KB requeriría la lectura de 128 bloques y el valor predeterminado de este parámetro sería, por lo tanto, de 128.

La mayoría de los problemas de rendimiento de la base de datos observados por NetApp en los sitios de los clientes implican una configuración incorrecta para este parámetro. Hay motivos válidos para cambiar este valor con las versiones 8 y 9 de Oracle. Como resultado, el parámetro puede estar presente sin saberlo en `init.ora` Archivos porque la base de datos se actualizó in situ a Oracle 10 y versiones posteriores. Una configuración heredada de 8 o 16, en comparación con el valor predeterminado de 128, daña significativamente el rendimiento de I/O secuencial.

NetApp recomienda configurar el `db_file_multiblock_read_count` el parámetro no debe estar presente en el `init.ora` archivo. NetApp nunca se ha encontrado con una situación en la que cambiar este parámetro mejoró el rendimiento, pero hay muchos casos en los que causó daños claros en el rendimiento de I/O secuencial.

filesystemio_options

Parámetro de inicialización de Oracle `filesystemio_options` Controla el uso de la E/S asíncrona y directa

El comportamiento y las recomendaciones para filesystemio_options en ASA r2 son idénticos a los sistemas AFF/ FAS porque el parámetro es específico de Oracle y no depende de la plataforma de almacenamiento. ASA r2 usa ONTAP como AFF/ FAS, por lo que se aplican las mismas prácticas recomendadas.

Contrariamente a la creencia común, las E/S asincrónicas y directas no son mutuamente excluyentes. NetApp ha observado que este parámetro suele estar mal configurado en los entornos del cliente, y esta mala configuración es el responsable directo de muchos problemas de rendimiento.

La E/S asíncrona significa que las operaciones de I/O de Oracle se pueden paralelizar. Antes de la disponibilidad de E/S asíncrona en varios sistemas operativos, los usuarios configuraron numerosos procesos de escritura de base de datos y cambiaron la configuración del proceso del servidor. Con la E/S asíncrona, el propio sistema operativo realiza E/S en nombre del software de base de datos de forma paralela y altamente eficiente. Este proceso no pone los datos en riesgo y las operaciones críticas, como el redo registro de Oracle, se siguen realizando de forma síncrona.

La E/S directa omite la caché de buffers del SO. Las E/S en un sistema UNIX normalmente fluyen a través de la caché de buffers del sistema operativo. Esto es útil para aplicaciones que no mantienen una caché interna, pero Oracle tiene su propia caché de buffers en SGA. En casi todos los casos, es mejor habilitar la E/S directa y asignar la RAM del servidor a la SGA en lugar de confiar en la caché de buffers del SO. Oracle SGA utiliza la memoria de forma más eficaz. Además, cuando la I/O fluye por el búfer del SO, se somete a un procesamiento adicional, lo que aumenta las latencias. El aumento de las latencias es especialmente notable en operaciones pesadas de I/O de escritura cuando un requisito crucial es la baja latencia.

Las opciones para `filesystemio_options` son:

- **Async.** Oracle envía solicitudes de E/S al sistema operativo para su procesamiento. Este proceso permite a Oracle realizar otro trabajo en lugar de esperar la finalización de E/S y, por lo tanto, aumenta la paralelización de E/S.
- **Directio.** Oracle realiza E/S directamente contra archivos físicos en lugar de enrutar E/S a través de la caché del SO host.
- **None.** Oracle utiliza E/S síncronas y en buffer En esta configuración, la elección entre los procesos de servidor compartido y dedicado y el número de dbwriters son más importantes.
- **Setall.** Oracle utiliza E/S tanto asíncrona como directa En casi todos los casos, el uso de `setall` es óptimo.

> En entornos ASM, Oracle utiliza automáticamente E/S directa y E/S asincrónica para discos administrados por ASM, por lo que `filesystemio_options` No tiene ningún efecto sobre los grupos de discos ASM. Para implementaciones que no sean ASM (por ejemplo, sistemas de archivos en LUN SAN), configure: `filesystemio_options = setall`. Esto permite E/S directa y asincrónica para un rendimiento óptimo.

Algunos sistemas operativos más antiguos tenían problemas con la E/S asincrónica, lo que dio lugar a recomendaciones obsoletas que sugerían evitarla. Sin embargo, la E/S asincrónica es estable y totalmente compatible con todos los sistemas operativos actuales. No hay razón para desactivarlo a menos que se identifique un error específico del sistema operativo.

Si una base de datos ha estado utilizando E/S en búfer, un cambio a E/S directa también puede justificar un

cambio en el tamaño de SGA. Al desactivar las E/S en buffer, se elimina la ventaja de rendimiento que proporciona la caché del SO del host para la base de datos. Al volver a agregar RAM a SGA se soluciona este problema. El resultado neto debe ser una mejora en el rendimiento de E/S.

Aunque casi siempre es mejor utilizar RAM para Oracle SGA que para el almacenamiento en caché de buffers del sistema operativo, puede ser imposible determinar el mejor valor. Por ejemplo, puede ser preferible utilizar E/S en buffer con tamaños SGA muy pequeños en un servidor de bases de datos con muchas instancias de Oracle activas de forma intermitente. Esta disposición permite el uso flexible de la RAM libre restante en el SO por todas las instancias de base de datos en ejecución. Se trata de una situación muy inusual, pero se ha observado en algunos sitios de clientes.

* NetApp recomienda* configuración `filesystemio_options` a `setall`, pero tenga en cuenta que, en algunas circunstancias, la pérdida de la memoria caché del búfer del host puede requerir un aumento en Oracle SGA. Los sistemas ASA r2 están optimizados para cargas de trabajo SAN con baja latencia, por lo que el uso de setall se alinea perfectamente con el diseño de ASA para implementaciones de Oracle de alto rendimiento.

Tiempo de espera de RAC

Oracle RAC es un producto de clusterware con varios tipos de procesos internos de latido que controlan el estado del cluster.

Los sistemas ASA r2 usan ONTAP al igual que AFF/ FAS, por lo que se aplican los mismos principios para los parámetros de tiempo de espera de Oracle RAC. No hay cambios específicos de ASA en las recomendaciones de tiempo de espera de disco o de recuento de errores. Sin embargo, ASA r2 está optimizado para cargas de trabajo SAN y conmutación por error de baja latencia, lo que hace que estas prácticas recomendadas sean aún más críticas.

La información contenida en el "recuento de errores" Esta sección incluye información crítica para entornos Oracle RAC que utilizan almacenamiento en red y, en muchos casos, será necesario cambiar la configuración predeterminada de Oracle RAC para garantizar que el clúster RAC sobreviva a los cambios de ruta de red y a las operaciones de conmutación por error de almacenamiento.

tiempo de espera del disco

El parámetro de RAC relacionado con el almacenamiento primario es `disktimeout`. Este parámetro controla el umbral en el que debe completarse la E/S del archivo de quorum. Si la `disktimeout` Se supera el parámetro y el nodo RAC se expulsa del clúster. El valor predeterminado de este parámetro es 200. Este valor debería ser suficiente para los procedimientos estándar de toma de control y devolución del almacenamiento.

NetApp recomienda probar exhaustivamente las configuraciones de RAC antes de colocarlas en producción, ya que existen muchos factores que afectan a una toma de control o al retorno primario. Además del tiempo necesario para que se complete la conmutación por error del almacenamiento, también se requiere más tiempo para que se propaguen los cambios del protocolo de control de agregación de enlaces (LACP). Además, el software multivía SAN debe detectar un tiempo de espera de I/O y volver a intentarlo en una ruta alternativa. Si una base de datos está extremadamente activa, se debe poner en cola una gran cantidad de E/S y volver a intentarlo antes de procesar la E/S del disco de quorum.

Si no se puede realizar una toma de control o una devolución del almacenamiento real, el efecto se puede simular mediante pruebas de extracción de cables en el servidor de bases de datos.

NetApp recomienda lo siguiente:

- Dejando el `disktimeout` parámetro con el valor predeterminado de 200.
- Pruebe siempre a fondo una configuración de RAC.

recuento de errores

La `misscount` Normalmente, el parámetro sólo afecta al latido de red entre los nodos de RAC. El valor predeterminado es 30 segundos. Si los binarios de grid se encuentran en una cabina de almacenamiento o si la unidad de arranque del sistema operativo no es local, este parámetro puede volverse importante. Esto incluye hosts con unidades de arranque ubicadas en una SAN FC, sistemas operativos arrancados NFS y unidades de arranque ubicadas en almacenes de datos de virtualización, como un archivo VMDK.

Si el acceso a una unidad de arranque se interrumpe por una toma de control o una restauración del almacenamiento, es posible que la ubicación binaria del grid o todo el sistema operativo se bloquee temporalmente. El tiempo necesario para que ONTAP complete la operación de almacenamiento y que el sistema operativo cambie de rutas y reanude las I/O puede superar el `misscount` umbral. Como resultado, un nodo se expulsa inmediatamente después de restaurar la conectividad con el LUN de arranque o los binarios de grid. En la mayoría de los casos, la expulsión y el reinicio posterior se producen sin mensajes de registro que indiquen el motivo del reinicio. No todas las configuraciones se ven afectadas, por lo que debe realizar pruebas de cualquier host basado en almacenes de datos, arranque en NFS o arranque en SAN en un entorno RAC para que RAC se mantenga estable si se interrumpe la comunicación con la unidad de arranque.

En el caso de unidades de arranque no locales o de un sistema de archivos no local `grid` binarios, el `misscount` será necesario cambiar para que coincida `disktimeout`. Si se cambia este parámetro, realice otras pruebas para identificar también cualquier efecto sobre el comportamiento de RAC, como el tiempo de conmutación por error del nodo.

NetApp recomienda lo siguiente:

- Abandone el `misscount` parámetro con el valor por defecto de 30 a menos que se aplique una de las siguientes condiciones:
  - `grid` Los archivos binarios se encuentran en una unidad conectada a la red, incluidas las unidades basadas en iSCSI, FC y almacén de datos.
  - El sistema operativo se inicia mediante SAN.
- En tales casos, evalúe el efecto de las interrupciones de la red que afectan el acceso al sistema operativo o. `GRID_HOME` sistemas de ficheros: En algunos casos, estas interrupciones provocan que los daemons de Oracle RAC se atasquen, lo que puede provocar un `misscount`-basado en tiempo de espera y desalojo. El tiempo de espera predeterminado es 27 segundos, que es el valor de `misscount` menos `reboottime`. En tales casos, aumentar `misscount` a 200 para coincidir `disktimeout`.

- El diseño optimizado para SAN de ASA r2 reduce la latencia de conmutación por error, pero los tiempos de espera aún deben ajustarse para los binarios de arranque en red o de cuadrícula.
- Para configuraciones RAC extendidas o activo-activo (por ejemplo, sincronización activa de SnapMirror), el ajuste del tiempo de espera sigue siendo esencial para las arquitecturas de RPO cero.

Configuración de host con sistemas AFF/ FAS

AIX

Temas de configuración para bases de datos de Oracle en IBM AIX con ONTAP.

I/O concurrente

Lograr un rendimiento óptimo en IBM AIX requiere el uso de E/S concurrentes Sin operaciones de I/O simultáneas, es probable que las limitaciones de rendimiento se deban a que AIX realiza I/O atómicas serializadas, lo que conlleva una sobrecarga significativa.

En un principio, NetApp recomendó utilizar el `cio` Opción de montaje para forzar el uso de E/S concurrentes en el sistema de archivos, pero este proceso tenía inconvenientes y ya no es necesario. Desde la introducción de AIX 5,2 y Oracle 10gR1, Oracle en AIX puede abrir archivos individuales para I/O simultánea, en lugar de forzar las operaciones de I/O simultáneas en todo el sistema de archivos.

El mejor método para habilitar E/S concurrente es establecer el `init.ora` parámetro `filesystemio_options` para `setall`. Al hacerlo, Oracle puede abrir archivos específicos para utilizarlos con E/S simultáneas

Uso `cio` Como opción de montaje fuerza el uso de I/O concurrente, lo cual puede tener consecuencias negativas. Por ejemplo, al forzar la E/S simultánea se desactiva la lectura anticipada en los sistemas de archivos, lo que puede dañar el rendimiento de las E/S que se producen fuera del software de la base de datos Oracle, como copiar archivos y realizar copias de seguridad en cinta. Además, productos como Oracle GoldenGate y SAP BR*Tools no son compatibles con el uso del `cio` Opción de montaje con determinadas versiones de Oracle.

NetApp recomienda lo siguiente:

- No utilice la `cio` opción de montaje en el nivel de sistema de archivos. En su lugar, habilite la I/O simultánea mediante el uso de `filesystemio_options=setall`.
- Utilice sólo el `cio` la opción de montaje debería si no es posible configurarla `filesystemio_options=setall`.

Opciones de montaje de AIX NFS

En la siguiente tabla, se enumeran las opciones de montaje de AIX NFS para bases de datos de instancia única de Oracle.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Archivos de control Archivos de datos Rehacer registros	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, intr`

En la siguiente tabla, se enumeran las opciones de montaje de AIX NFS para RAC.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Archivos de control Archivos de datos Rehacer registros	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac`
`CRS/Voting`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac`
Específico `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Compartido `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

La diferencia principal entre las opciones de montaje de instancia única y RAC es la adición de `noac` a las opciones de montaje. Esta adición tiene el efecto de deshabilitar el almacenamiento en caché del SO del host que permite que todas las instancias del clúster RAC tengan una vista uniforme del estado de los datos.

Aunque utilice el `cio` monte la opción y la `init.ora` parámetro `filesystemio_options=setall` tiene el mismo efecto de deshabilitar el almacenamiento en caché de host, sigue siendo necesario utilizarlo `noac`. `noac` es necesario para el uso compartido `ORACLE_HOME` Despliegues para facilitar la coherencia de archivos como archivos de contraseñas de Oracle y. `spfile` archivos de parámetros. Si cada instancia de un clúster de RAC tiene un dedicado `ORACLE_HOME`, entonces este parámetro no es necesario.

Opciones de montaje jfs/JFS2 de AIX

En la siguiente tabla se enumeran las opciones de montaje jfs/JFS2 de AIX.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	Valores predeterminados
Archivos de control Archivos de datos Rehacer registros	Valores predeterminados
ORACLE_HOME	Valores predeterminados

Antes de utilizar AIX `hdisk` los dispositivos de cualquier entorno, incluidas las bases de datos, comprueban el parámetro `queue_depth`. Este parámetro no es la profundidad de la cola del HBA, más bien se relaciona con la profundidad de la cola SCSI de una persona `hdisk device. Depending on how the LUNs are configured, the value for `queue_depth` puede ser demasiado bajo para un buen rendimiento. Las pruebas han demostrado que el valor óptimo es 64.

HP-UX

Temas de configuración para bases de datos de Oracle en HP-UX con ONTAP.

Opciones de montaje NFS de HP-UX

En la siguiente tabla se enumeran las opciones de montaje de HP-UX NFS para una única instancia.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Archivos de control Archivos de datos Rehacer registros	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144,forc edirectio, nointr,suid`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`

En la siguiente tabla se enumeran las opciones de montaje de HP-UX NFS para RAC.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac,suid`
Archivos de control Archivos de datos Rehacer registros	`rw, bg,hard, [vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio,suid`
CRS/votación	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac, forcedirectio,suid`
Específico `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Compartido `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,suid`

La diferencia principal entre las opciones de montaje de instancia única y RAC es la adición de `noac` y.. `forcedirectio` a las opciones de montaje. Esta adición tiene el efecto de deshabilitar el almacenamiento en caché del sistema operativo del host, lo que permite que todas las instancias del clúster RAC tengan una vista coherente del estado de los datos. Aunque utilice el `init.ora` parámetro `filesystemio_options=setall` tiene el mismo efecto de deshabilitar el almacenamiento en caché de host, sigue siendo necesario utilizarlo `noac` y.. `forcedirectio`.

La razón `noac` es necesario para el uso compartido `ORACLE_HOME` Despliegues es para facilitar la coherencia de archivos como archivos de contraseñas de Oracle y archivos spfiles. Si cada instancia de un clúster de RAC tiene un dedicado `ORACLE_HOME`, este parámetro no es necesario.

Opciones de montaje HP-UX VxFS

Utilice las siguientes opciones de montaje para sistemas de archivos que alojan binarios de Oracle:

```
delaylog,nodatainlog
```

Utilice las siguientes opciones de montaje para sistemas de archivos que contienen archivos de datos, redo logs, archive logs y archivos de control en los que la versión de HP-UX no admite E/S simultáneas:

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

Cuando se admiten E/S simultáneas (VxFS 5.0.1 y posteriores, o con ServiceGuard Storage Management Suite), utilice estas opciones de montaje para sistemas de archivos que contengan archivos de datos, redo logs, archive logs y archivos de control:

```
delaylog,cio
```

> El parámetro `db_file_multiblock_read_count` Es especialmente crítico en entornos VxFS. Oracle recomienda que este parámetro permanezca sin definir en Oracle 10g R1 y posteriores a menos que se indique lo contrario específicamente. El valor por defecto con un tamaño de bloque de Oracle 8KB es 128. Si el valor de este parámetro se fuerza a 16 o menos, quite el `convosync=direct` Opción de montaje porque puede dañar el rendimiento de I/O secuencial. Este paso daña otros aspectos del rendimiento y solo debe tomarse si el valor de `db_file_multiblock_read_count` debe cambiarse a partir del valor predeterminado.

Linux

Temas de configuración específicos del sistema operativo Linux.

Tablas de ranuras TCP Linux NFSv3

Las tablas de ranuras TCP son equivalentes a NFSv3 a la profundidad de la cola del adaptador de bus de host (HBA). En estas tablas se controla el número de operaciones de NFS que pueden extraordinarias a la vez. El valor predeterminado suele ser 16, que es demasiado bajo para un rendimiento óptimo. El problema opuesto ocurre en los kernels más nuevos de Linux, que pueden aumentar automáticamente el límite de la tabla de ranuras TCP a un nivel que sature el servidor NFS con solicitudes.

Para obtener un rendimiento óptimo y evitar problemas de rendimiento, ajuste los parámetros del núcleo que controlan las tablas de ranuras TCP.

Ejecute el `sysctl -a | grep tcp.*.slot_table` command, y observe los siguientes parámetros:

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

Todos los sistemas Linux deben incluir `sunrpc.tcp_slot_table_entries`, pero solo algunos incluyen `sunrpc.tcp_max_slot_table_entries`. Ambos deben establecerse en 128.

> Si no se establecen estos parámetros, puede tener efectos significativos en el rendimiento. En algunos casos, el rendimiento es limitado porque el sistema operativo linux no está emitiendo suficiente I/O. En otros casos, las latencias de I/O aumentan cuando el sistema operativo linux intenta emitir más operaciones de I/O de las que se pueden mantener.

Opciones de montaje de Linux NFS

En la siguiente tabla, se enumeran las opciones de montaje de NFS de Linux para una instancia única.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Archivos de control Archivos de datos Rehacer registros	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`
ORACLE_HOME	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

La siguiente tabla enumera las opciones de montaje de NFS de Linux para RAC.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, actimeo=0`
Archivos de control Archivos de datos Rehacer registros	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`
CRS/votación	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,actimeo=0`
Específico `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Compartido `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`

La diferencia principal entre las opciones de montaje de instancia única y RAC es la adición de `actimeo=0` a las opciones de montaje. Esta adición tiene el efecto de deshabilitar el almacenamiento en caché del sistema operativo del host, lo que permite que todas las instancias del clúster RAC tengan una vista coherente del estado de los datos. Aunque utilice el `init.ora` parámetro `filesystemio_options=setall` tiene el mismo efecto de deshabilitar el almacenamiento en caché de host, sigue siendo necesario utilizarlo `actimeo=0`.

La razón `actimeo=0` es necesario para el uso compartido `ORACLE_HOME` Despliegues es para facilitar la consistencia de archivos como los archivos de contraseñas de Oracle y spfiles. Si cada instancia de un clúster de RAC tiene un dedicado `ORACLE_HOME`, entonces este parámetro no es necesario.

Por lo general, los archivos que no son de base de datos se deben montar con las mismas opciones utilizadas para los archivos de datos de instancia única, aunque las aplicaciones específicas pueden tener requisitos diferentes. Evite las opciones de montaje `noac` y.. `actimeo=0` si es posible, ya que estas opciones desactivan la lectura anticipada y el almacenamiento en búfer de nivel de sistema de archivos. Esto puede causar graves problemas de rendimiento en procesos como extracción, traducción y carga.

ACCESO y GETATTR

Algunos clientes han observado que un nivel extremadamente alto de otros IOPS, como EL ACCESO y GETATTR, puede dominar sus cargas de trabajo. En casos extremos, las operaciones como las de lectura y escritura pueden ser tan bajas como el 10 % del total. Este es un comportamiento normal con cualquier base de datos que incluya el uso `actimeo=0` y/o. `noac` En Linux, porque estas opciones provocan que el sistema operativo Linux vuelva a cargar constantemente los metadatos de los archivos del sistema de almacenamiento. Las operaciones como EL ACCESO y GETATTR son operaciones de bajo impacto que se proporcionan desde la caché de ONTAP en un entorno de base de datos. No se deben considerar IOPS auténticos, como lecturas y escrituras, que crean una demanda real de sistemas de almacenamiento. Sin embargo, estas otras IOPS crean una cierta carga, sobre todo en entornos RAC. Para solucionar esta situación, habilite DNFS, que omite la caché de buffers del sistema operativo y evita estas operaciones de metadatos innecesarias.

NFS directo de Linux

Una opción de montaje adicional denominada `nosharecache`, Es necesario cuando (a) DNFS está activado y (b) un volumen de origen se monta más de una vez en un único servidor (c) con un montaje NFS anidado. Esta configuración se ve principalmente en entornos que admiten aplicaciones SAP. Por ejemplo, un único volumen de un sistema NetApp podría tener un directorio ubicado en `/vol/oracle/base` y un segundo en `/vol/oracle/home`. Si `/vol/oracle/base` está montado en `/oracle` y.. `/vol/oracle/home` está montado en `/oracle/home`, El resultado son montajes NFS anidados que se originan en la misma fuente.

El sistema operativo puede detectar el hecho de que `/oracle` y `/oracle/home` residen en el mismo volumen, que es el mismo sistema de archivos de origen. A continuación, el sistema operativo utiliza el mismo identificador de dispositivo para acceder a los datos. Al hacerlo, se mejora el uso del almacenamiento en caché del sistema operativo y algunas otras operaciones, pero interfiere con DNFS. Si DNFS debe acceder a un archivo, como `spfile` , activado `/oracle/home`, puede intentar utilizar erróneamente la ruta de acceso incorrecta a los datos. El resultado es una operación de I/O con errores. En estas configuraciones, agregue `nosharecache` la opción de montaje a cualquier sistema de archivos NFS que comparta un volumen de origen con otro sistema de archivos NFS en ese host. Al hacerlo, se fuerza al sistema operativo Linux a asignar un identificador de dispositivo independiente para ese sistema de archivos.

Linux Direct NFS y Oracle RAC

El uso de DNFS ofrece ventajas especiales de rendimiento para Oracle RAC en el sistema operativo Linux, ya que Linux no dispone de un método para forzar la entrada/salida directa, que se necesita con RAC para lograr coherencia entre los nodos. Como solución alternativa, Linux requiere el uso de `actimeo=0` Opción de montaje, que hace que los datos de archivo caduquen inmediatamente desde la caché del sistema operativo. Esta opción, a su vez, fuerza al cliente NFS de Linux a volver a leer constantemente los datos de atributos, lo que daña la latencia y aumenta la carga en la controladora de almacenamiento.

Al habilitar DNFS se omite el cliente NFS del host y se evita este daño. Varios clientes han informado de mejoras significativas en el rendimiento en clústeres RAC y reducciones considerables en la carga de ONTAP

(especialmente con respecto a otras IOPS) al habilitar DNFS.

Linux Direct NFS y archivo oranfstab

Al utilizar DNFS en Linux con la opción multipathing, se deben utilizar varias subredes. En otros sistemas operativos, se pueden establecer varios canales DNFS mediante el `LOCAL` y.. `DONTROUTE` Opciones para configurar varios canales DNFS en una sola subred. Sin embargo, esto no funciona correctamente en Linux y puede resultar en problemas de rendimiento inesperados. Con Linux, cada NIC utilizada para el tráfico DNFS debe estar en una subred diferente.

Programador de I/O.

El kernel de Linux permite un control de bajo nivel sobre la forma en que se programa la E/S para bloquear los dispositivos. Los valores por defecto en varias distribuciones de Linux varían considerablemente. Las pruebas demuestran que la fecha límite suele ofrecer los mejores resultados, pero en ocasiones NOOP ha sido ligeramente mejor. La diferencia de rendimiento es mínima, pero pruebe ambas opciones si es necesario extraer el máximo rendimiento posible de una configuración de base de datos. CFQ es el valor predeterminado en muchas configuraciones y ha demostrado tener problemas de rendimiento significativos con cargas de trabajo de bases de datos.

Consulte la documentación relevante del proveedor de Linux para obtener instrucciones sobre la configuración del programador de E/S.

Accesos múltiples

Algunos clientes se han encontrado con fallos durante la interrupción de la red porque el daemon multivía no se estaba ejecutando en su sistema. En versiones recientes de Linux, el proceso de instalación del sistema operativo y el daemon de rutas múltiples pueden dejar estos sistemas operativos vulnerables a este problema. Los paquetes están instalados correctamente, pero no están configurados para el inicio automático después de un reinicio.

Por ejemplo, el valor predeterminado para el daemon multipath en RHEL5,5 puede aparecer del siguiente modo:

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

Esto se puede corregir con los siguientes comandos:

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

Duplicación de ASM

La duplicación de ASM puede requerir cambios en la configuración multivía de Linux para permitir que ASM reconozca un problema y cambie a un grupo de fallos alternativo. La mayoría de las configuraciones de ASM en ONTAP utilizan redundancia externa, lo que significa que la cabina externa ofrece protección de datos y ASM no refleja datos. Algunos sitios utilizan ASM con redundancia normal para proporcionar duplicación bidireccional, normalmente en diferentes sitios.

La configuración de Linux que se muestra en la "Documentación de utilidades de host de NetApp" Incluya parámetros multivía que generen la cola indefinida de I/O. Esto significa que una I/O en un dispositivo LUN sin rutas activas espera tanto tiempo como sea necesario para que finalice la I/O. Esto suele ser deseable ya que los hosts Linux esperan todo el tiempo necesario para que se completen los cambios de ruta SAN, para que se reinicien los switches FC o para que un sistema de almacenamiento complete una conmutación al respaldo.

Este comportamiento de puesta en cola ilimitada provoca un problema con el mirroring de ASM debido a que ASM debe recibir un error de I/O para que vuelva a intentar I/O en un LUN alternativo.

Defina los siguientes parámetros en Linux `multipath.conf` Archivo para LUN de ASM utilizados con la duplicación de ASM:

```
polling_interval 5
no_path_retry 24
```

Estos valores crean un timeout de 120 segundos para los dispositivos ASM. El tiempo de espera se calcula como el `polling_interval` * `no_path_retry` como segundos. Puede que sea necesario ajustar el valor exacto en algunas circunstancias, pero un tiempo de espera de 120 segundos debería ser suficiente para la mayoría de los usos. Concretamente, 120 segundos deberían permitir que se produzca una toma de control o una devolución de la controladora sin que se produzca un error de I/O, lo que provocaría que el grupo de errores se desconectara.

A inferior `no_path_retry` Value puede reducir el tiempo necesario para que ASM cambie a un grupo de fallos alternativo, pero esto también aumenta el riesgo de una conmutación por error no deseada durante actividades de mantenimiento como la toma de control de un controlador. El riesgo se puede mitigar mediante una supervisión cuidadosa del estado de duplicación de ASM. Si se produce una conmutación al respaldo no deseada, los duplicados pueden volver a sincronizarse rápidamente si la resincronización se realiza con relativa rapidez. Para obtener información adicional, consulte la documentación de Oracle on ASM Fast Mirror Resync para ver la versión del software de Oracle en uso.

Opciones de montaje de Linux xfs, ext3 y ext4

NetApp recomienda usar las opciones de montaje predeterminadas.

ASMLib/AFD (Controlador de Filtro de ASM)

Temas de configuración específicos del sistema operativo Linux mediante AFD y ASMLib

Tamaños de bloque ASMLib

ASMLib es una biblioteca de gestión de ASM opcional y utilidades asociadas. Su valor principal es la capacidad para estampar un LUN o un archivo basado en NFS como un recurso ASM con una etiqueta legible para el ser humano.

Las versiones recientes de ASMLib detectan un parámetro de LUN llamado Logical Blocks per Physical Block Exponent (LBPPBE). El destino SCSI de ONTAP no notificó este valor hasta hace poco. Ahora devuelve un valor que indica que se prefiere un tamaño de bloque de 4KB KB. Esta no es una definición de tamaño de bloque, pero es una indicación para cualquier aplicación que utilice LBPPBE de que las E/S de un determinado tamaño podrían manejarse de manera más eficiente. Sin embargo, ASMLib interpreta LBPPBE como un tamaño de bloque y marca de forma persistente la cabecera ASM cuando se crea el dispositivo ASM.

Este proceso puede causar problemas con actualizaciones y migraciones de varias maneras, todo ello en función de la incapacidad de mezclar dispositivos ASMLib con diferentes tamaños de bloque en el mismo grupo de discos ASM.

Por ejemplo, las matrices más antiguas generalmente reportaron un valor LBPPBE de 0 o no reportaron este valor en absoluto. ASMLib lo interpreta como un tamaño de bloque de 512 bytes. Las cabinas más recientes se interpretarán con un tamaño de bloque de 4KB KB. No es posible mezclar dispositivos de 512 bytes y 4KB en el mismo grupo de discos ASM. Al hacerlo, se bloquearía a un usuario para que no aumente el tamaño del grupo de discos de ASM utilizando LUN de dos matrices o aprovechando ASM como herramienta de migración. En otros casos, es posible que RMAN no permita la copia de archivos entre un grupo de discos de ASM con un tamaño de bloque de 512 bytes y un grupo de discos de ASM con un tamaño de bloque de 4KB KB.

La solución preferida es parchear ASMLib. El identificador de error de Oracle es 13999609 y el parche está presente en oracleasm-support-2,1.8-1 y superior. Este parche permite al usuario definir el parámetro `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` para `true` en la `/etc/sysconfig/oracleasm` archivo de configuración. Al hacerlo, se bloquea ASMLib para que no utilice el parámetro LBPPBE, lo que significa que los LUN de la nueva matriz ahora se reconocen como dispositivos de bloque de 512 bytes.

> La opción no cambia el tamaño de bloque en LUN que ASMLib estampó anteriormente. Por ejemplo, si un grupo de discos ASM con bloques de 512 bytes debe migrarse a un nuevo sistema de almacenamiento que notifique un bloque de 4KB KB, la opción `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Debe establecerse antes de que las nuevas LUN se estampen con ASMLib. Si los dispositivos ya han sido estampados por oracleasm, deben ser reformateados antes de ser reincorporados con un nuevo tamaño de bloque. En primer lugar, desconfigure el dispositivo con `oracleasm deletedisk`, Y, a continuación, borre los primeros 1GB del dispositivo con `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Por último, si el dispositivo se ha particionado previamente, utilice `kpartx` Comando para eliminar particiones obsoletas o simplemente reiniciar el sistema operativo.

Si no se puede aplicar un parche a ASMLib, se puede eliminar ASMLib de la configuración. Este cambio es disruptivo y requiere el desestampado de discos de ASM y asegurarse de que el `asm_diskstring` el parámetro se ha definido correctamente. Sin embargo, este cambio no requiere la migración de datos.

Tamaños de bloque de unidad de filtro de ASM (AFD)

AFD es una biblioteca de gestión de ASM opcional que se está convirtiendo en el reemplazo de ASMLib. Desde el punto de vista del almacenamiento, es muy similar a ASMLib, pero incluye características adicionales como la capacidad de bloquear E/S no Oracle para reducir las posibilidades de errores de usuario o aplicación que podrían dañar los datos.

Tamaños de bloques de dispositivos

Al igual que ASMLib, AFD también lee el parámetro LUN Bloques lógicos por Exponente de bloque físico (LBPPBE) y utiliza de forma predeterminada el tamaño del bloque físico, no el tamaño del bloque lógico.

Esto podría crear un problema si se agrega AFD a una configuración existente donde los dispositivos ASM ya están formateados como dispositivos de bloque de 512 bytes. El controlador AFD reconocería el LUN como un dispositivo 4K y la discrepancia entre la etiqueta ASM y el dispositivo físico impediría el acceso. Del mismo modo, las migraciones se verían afectadas porque no es posible mezclar dispositivos de 512 bytes y 4KB en el mismo grupo de discos de ASM. Al hacerlo, se bloquearía a un usuario para que no aumente el tamaño del grupo de discos de ASM utilizando LUN de dos matrices o aprovechando ASM como herramienta de migración. En otros casos, es posible que RMAN no permita la copia de archivos entre un grupo de discos de ASM con un tamaño de bloque de 512 bytes y un grupo de discos de ASM con un tamaño de bloque de 4KB

KB.

La solución es simple: AFD incluye un parámetro para controlar si utiliza los tamaños de bloque lógicos o físicos. Este es un parámetro global que afecta a todos los dispositivos del sistema. Para forzar a AFD a utilizar el tamaño de bloque lógico, establezca `options oracleafd oracleafd_use_logical_block_size=1` en la `/etc/modprobe.d/oracleafd.conf` archivo.

Tamaños de transferencia multivía

Los cambios recientes del kernel de linux aplican las restricciones de tamaño de I/O enviadas a dispositivos multivía y el AFD no cumple con estas restricciones. A continuación, se rechazan las I/O, lo que hace que la ruta de LUN se desconecte. El resultado es una incapacidad para instalar Oracle Grid, configurar ASM o crear una base de datos.

La solución es especificar manualmente la longitud máxima de transferencia del archivo multipath.conf para las LUN de ONTAP:

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

Incluso si no existe ningún problema en la actualidad, este parámetro debe configurarse si se utiliza AFD para garantizar que una actualización de linux futura no cause problemas de forma inesperada.

Microsoft Windows

Temas de configuración de bases de datos de Oracle en Microsoft Windows con ONTAP.

NFS

Oracle admite el uso de Microsoft Windows con el cliente NFS directo. Esta funcionalidad ofrece un acceso a las ventajas de gestión de NFS, incluida la capacidad de ver archivos de distintos entornos, cambiar el tamaño de volúmenes de forma dinámica y utilizar un protocolo IP menos costoso. Consulte la documentación oficial de Oracle para obtener información sobre la instalación y configuración de una base de datos en Microsoft Windows mediante DNFS. No existen mejores prácticas especiales.

SAN

Para una eficiencia de compresión óptima, asegúrese de que el sistema de archivos NTFS utilice una unidad de asignación de 8K o más. El uso de una unidad de asignación de 4K, que suele ser la predeterminada, afecta negativamente a la eficiencia de la compresión.

Solaris

Temas de configuración específicos del sistema operativo Solaris.

Opciones de montaje NFS de Solaris

En la siguiente tabla se enumeran las opciones de montaje NFS de Solaris para una única instancia.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	`rw,bg,hard,[vers=3,vers=4.1], roto=tcp, timeo=600,rsize=262144,wsize=262144`
Archivos de control Archivos de datos Rehacer registros	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,llock,suid`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`

Uso de `llock` se ha demostrado que mejora drásticamente el rendimiento en entornos de cliente al eliminar la latencia asociada con la adquisición y la liberación de bloqueos en el sistema de almacenamiento. Utilice esta opción con cuidado en entornos en los que se han configurado varios servidores para montar los mismos sistemas de archivos y Oracle está configurado para montar estas bases de datos. Aunque esta es una configuración muy inusual, es utilizada por un pequeño número de clientes. Si una instancia se inicia accidentalmente por segunda vez, se pueden producir daños en los datos porque Oracle no puede detectar los archivos de bloqueo en el servidor externo. Los bloqueos NFS no ofrecen protección de otro modo; al igual que en la versión 3 de NFS, solo son orientativos.

Debido a que el `llock` y.. `forcedirectio` los parámetros se excluyen entre sí, es importante hacerlo `filesystemio_options=setall` está presente en la `init.ora` archiva así `directio` se utiliza. Sin este parámetro, se utiliza el almacenamiento en caché del búfer del sistema operativo del host y el rendimiento se puede ver afectado negativamente.

En la siguiente tabla se muestran las opciones de montaje de Solaris NFS RAC.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac`
Archivos de control Archivos de datos Rehacer registros	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio`
CRS/votación	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio`
Específico `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Compartido `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,suid`

La diferencia principal entre las opciones de montaje de instancia única y RAC es la adición de `noac` y..

`forcedirectio` a las opciones de montaje. Esta adición tiene el efecto de deshabilitar el almacenamiento en caché del sistema operativo del host, lo que permite que todas las instancias del clúster RAC tengan una vista coherente del estado de los datos. Aunque utilice el `init.ora` parámetro `filesystemio_options=setall` tiene el mismo efecto de deshabilitar el almacenamiento en caché de host, sigue siendo necesario utilizarlo `noac` y.. `forcedirectio`.

La razón `actimeo=0` es necesario para el uso compartido `ORACLE_HOME` Despliegues es para facilitar la coherencia de archivos como archivos de contraseñas de Oracle y archivos spfiles. Si cada instancia de un clúster de RAC tiene un dedicado `ORACLE_HOME`, este parámetro no es necesario.

Opciones de montaje UFS de Solaris

NetApp recomienda usar la opción de montaje de registro para conservar la integridad de los datos en caso de bloqueo del host Solaris o interrupción de la conectividad de FC. La opción de montaje logging también conserva la facilidad de uso de los backups de Snapshot.

ZFS de Solaris

Solaris ZFS debe instalarse y configurarse cuidadosamente para ofrecer un rendimiento óptimo.

mvector

Solaris 11 incluyó un cambio en la forma en que procesa grandes operaciones de E/S, lo que puede dar lugar a graves problemas de rendimiento en las matrices de almacenamiento SAN. El problema está documentado en el informe de errores de seguimiento de NetApp 630173, que indica que Solaris 11 ZFS Performance Regression.

Esto no es un bug de ONTAP. Se trata de un defecto de Solaris cuyo seguimiento se realiza bajo los defectos de Solaris 7199305 y 7082975.

Puede consultar a los Servicios de Soporte Oracle para averiguar si la versión de Solaris 11 está afectada o puede probar la solución provisional cambiando `zfs_mvector_max_size` a un valor menor.

Para ello, ejecute el siguiente comando como root:

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

Si surge algún problema inesperado de este cambio, se puede revertir fácilmente ejecutando el siguiente comando como root:

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

Kernel

El rendimiento fiable de ZFS requiere un kernel de Solaris parcheado contra problemas de alineación de LUN. La corrección se introdujo con el parche 147440-19 en Solaris 10 y con SRU 10,5 para Solaris 11. Utilice sólo Solaris 10 y versiones posteriores con ZFS.

Configuración de LUN

Para configurar una LUN, complete los siguientes pasos:

1. Cree una LUN del tipo `solaris`.
2. Instale el kit de utilidades de host (HUK) adecuado especificado por el "Herramienta de matriz de interoperabilidad de NetApp (IMT)".
3. Siga las instrucciones del HUK exactamente como se describe. Los pasos básicos se describen a continuación, pero consulte la "documentación más reciente" para el procedimiento adecuado.
   a. Ejecute el `host_config` utilidad para actualizar el `sd.conf/sdd.conf` archivo. Al hacerlo, las unidades SCSI pueden detectar correctamente LUN de ONTAP.
   b. Siga las instrucciones proporcionadas por el `host_config` Utilidad para habilitar la entrada/salida multivía (MPIO).
   c. Reiniciar. Este paso es necesario para que cualquier cambio se reconozca en el sistema.
4. Cree particiones en las LUN y compruebe que están correctamente alineadas. Consulte el Apéndice B: Verificación de alineación de WAFL para obtener instrucciones sobre cómo probar y confirmar la alineación directamente.

zpools

Sólo se debe crear un zpool después de los pasos del "Configuración de LUN" se realizan. Si el procedimiento no se realiza correctamente, puede provocar una degradación grave del rendimiento debido a la alineación de E/S. Para un rendimiento óptimo en ONTAP es necesario alinear el I/O con un límite de 4K GbE en una unidad. Los sistemas de archivos creados en un zpool utilizan un tamaño de bloque efectivo que se controla mediante un parámetro denominado `ashift`, que se puede ver ejecutando el comando `zdb -C`.

Valor de `ashift` el valor por defecto es 9, que significa 2^9, o 512 bytes. Para un rendimiento óptimo, el `ashift` El valor debe ser 12 (2^12=4K). Este valor se define en el momento en que se crea zpool y no se puede cambiar, lo que significa que los datos en zpools con `ashift` los datos que no sean 12 se deben migrar copiando a un zpool recién creado.

Después de crear un zpool, verifique el valor de `ashift` antes de continuar. Si el valor no es 12, las LUN no se detectaron correctamente. Destruya zpool, verifique que todos los pasos mostrados en la documentación de utilidades de host relevantes se hayan realizado correctamente y vuelva a crear zpool.

Zpools y LDOMs de Solaris

Los LDOMs de Solaris crean un requisito adicional para asegurarse de que la alineación de E/S es correcta. Aunque un LUN se puede detectar correctamente como dispositivo 4K, un dispositivo virtual vdsk en un LDOM no hereda la configuración del dominio de E/S. El vdsk basado en esa LUN vuelve a tener de forma predeterminada un bloque de 512 bytes.

Se necesita un archivo de configuración adicional. En primer lugar, se deben aplicar parches a los LDOM individuales para el bug de Oracle 15824910 para activar las opciones de configuración adicionales. Este parche se ha portado a todas las versiones utilizadas actualmente de Solaris. Una vez que se aplica el parche a LDOM, está listo para la configuración de las nuevas LUN correctamente alineadas de la siguiente manera:

1. Identifique los LUN o LUN que se van a utilizar en el nuevo zpool. En este ejemplo, es el dispositivo c2d1.

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

2. Recuperar la instancia vdc de los dispositivos que se van a utilizar para una agrupación ZFS:

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

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

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

Esto significa que a la instancia de dispositivo 1 se le asigna un tamaño de bloque de 4096.

Como ejemplo adicional, supongamos que las instancias de vdsk 1 a 6 deben configurarse para un tamaño de bloque de 4K KB y. `/etc/path_to_inst` se lee de la siguiente manera:

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

4. La final `vdc.conf` el archivo debe contener lo siguiente:

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

Precaución
El LDOM debe reiniciarse después de configurar vdc.conf y crear vdsk. Este paso no se puede evitar. El cambio de tamaño del bloque solo se aplica después de un reinicio. Continúe con la configuración de zpool y asegúrese de que el ashift está correctamente ajustado en 12 como se ha descrito anteriormente.

Registro de Intención de ZFS (ZIL)

Por lo general, no hay razón para localizar el registro de intención ZFS (ZIL) en un dispositivo diferente. El registro puede compartir espacio con el pool principal. El uso principal de un ZIL separado es cuando se utilizan unidades físicas que carecen de las funciones de almacenamiento en caché de escritura en cabinas de almacenamiento modernas.

sesgo logarítmico

Ajuste la `logbias` Parámetro en sistemas de archivos ZFS que alojan datos de Oracle.

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

Usar este parámetro reduce los niveles generales de escritura. En los valores predeterminados, los datos escritos se confirman primero en el ZIL y, a continuación, en el pool de almacenamiento principal. Este enfoque es adecuado para una configuración que utiliza una configuración de unidad simple, que incluye un dispositivo ZIL basado en SSD y medios giratorios para el pool de almacenamiento principal. Esto se debe a que permite un commit en una sola transacción de I/O en el medio de menor latencia disponible.

Cuando se utiliza una cabina de almacenamiento moderna que incluye su propia funcionalidad de almacenamiento en caché, este método no suele ser necesario. En raras ocasiones, es posible que sea conveniente comprometer una escritura con una sola transacción en el registro, como una carga de trabajo que consta de escrituras aleatorias altamente concentradas y sensibles a la latencia. Existen consecuencias en la amplificación de escritura, ya que los datos registrados se escriben finalmente en el pool de almacenamiento principal, lo que provoca el doble de la actividad de escritura.

E/S directa

Muchas aplicaciones, incluidos los productos de Oracle, pueden omitir la caché de buffers del host activando la E/S directa Esta estrategia no funciona como se esperaba con los sistemas de archivos ZFS. Aunque se omite la caché de buffers del host, ZFS continúa almacenando los datos en caché. Esta acción puede

provocar resultados engañosos cuando se usan herramientas como fio o sio para realizar pruebas de rendimiento, ya que es difícil predecir si I/O está llegando al sistema de almacenamiento o si se está almacenando en caché localmente dentro del sistema operativo. Esta acción también hace que sea muy difícil utilizar estas pruebas sintéticas para comparar el rendimiento de ZFS con otros sistemas de archivos. Como cuestión práctica, hay poca o ninguna diferencia en el rendimiento del sistema de archivos con las cargas de trabajo de los usuarios reales.

Varios zpools

Las copias de seguridad basadas en instantáneas, las restauraciones, los clones y el archivado de datos basados en ZFS se deben realizar en el nivel de zpool y, por lo general, requieren varios zpools. Un zpool es análogo a un grupo de discos LVM y debe configurarse usando las mismas reglas. Por ejemplo, es probable que una base de datos se disponga mejor con los archivos de datos en los que reside `zpool1` y los registros de archivo, los archivos de control y los registros de recuperación en los que residen `zpool2`. Este enfoque permite realizar un backup dinámico estándar en el que la base de datos se coloca en modo de backup dinámico, seguido de una copia Snapshot de `zpool1`. A continuación, la base de datos se elimina del modo de backup dinámico, se fuerza el archivo de registro y una copia de Snapshot de `zpool2` se ha creado. Una operación de restauración requiere el desmontaje de los sistemas de archivos zfs y desconectar zpool íntegramente, a continuación de una operación de restauración de SnapRestore. El zpool se puede poner en línea de nuevo y la base de datos se recupera.

filesystemio_options

Parámetro de Oracle `filesystemio_options` Funciona de forma diferente con ZFS. Si `setall` o. `directio` Se utiliza, las operaciones de escritura son síncronas y omiten la caché de buffers del sistema operativo, pero ZFS almacena en búfer las lecturas. Esta acción causa dificultades en el análisis de rendimiento porque a veces la caché ZFS intercepta y suministra servicio a las E/S, lo que hace que la latencia de almacenamiento y el total de E/S sean menores de lo que podría parecer.

Configuración de host con sistemas ASA r2

AIX

Temas de configuración para la base de datos Oracle en IBM AIX con ASA r2 ONTAP.

AIX es compatible con NetApp ASA r2 para alojar bases de datos Oracle, siempre que:

- Configura Oracle correctamente para E/S simultánea.
- Utiliza protocolos SAN compatibles (FC/iSCSI/NVMe).
- Ejecuta ONTAP 9.16.x o posterior en ASA r2.

I/O concurrente

Para lograr un rendimiento óptimo en IBM AIX con ASA r2 se requiere el uso de E/S simultánea. Sin E/S concurrente, es probable que haya limitaciones de rendimiento porque AIX realiza E/S atómica y serializada, lo que genera una sobrecarga significativa.

Originalmente, NetApp recomendaba utilizar el `cio` Opción de montaje para forzar E/S simultánea en el sistema de archivos, pero este proceso tenía inconvenientes y ya no es necesario. Desde la introducción de AIX 5.2 y Oracle 10gR1, Oracle en AIX puede abrir archivos individuales para E/S simultáneas, en lugar de forzar la E/S simultánea en todo el sistema de archivos.

El mejor método para habilitar E/S concurrente es establecer el `init.ora` parámetro `filesystemio_options` para `setall`. Al hacerlo, Oracle puede abrir archivos específicos para utilizarlos con E/S simultáneas

El uso de cio como opción de montaje fuerza el uso de E/S simultánea, lo que puede tener consecuencias negativas. Por ejemplo, forzar E/S simultáneas deshabilita la lectura anticipada en los sistemas de archivos, lo que puede dañar el rendimiento de las E/S que ocurren fuera del software de base de datos de Oracle, como copiar archivos y realizar copias de seguridad en cinta. Además, productos como Oracle GoldenGate y SAP BR*Tools no son compatibles con el uso de la opción de montaje cio con ciertas versiones de Oracle.

NetApp recomienda lo siguiente:

- No utilice la `cio` opción de montaje en el nivel de sistema de archivos. En su lugar, habilite la I/O simultánea mediante el uso de `filesystemio_options=setall`.
- Utilice únicamente el `cio` Opción de montaje si no es posible configurarla `filesystemio_options=setall`.

Dado que ASA r2 no es compatible con NAS, todas las implementaciones de Oracle en AIX deben utilizar protocolos de bloque.

Opciones de montaje jfs/JFS2 de AIX

En la siguiente tabla se enumeran las opciones de montaje jfs/JFS2 de AIX.

Tipo de archivo	Opciones de montaje
Directorio Raíz de ADR	Valores predeterminados
Archivos de control	Valores predeterminados
Archivos de datos	Valores predeterminados
Registros de rehacer	Valores predeterminados
ORACLE_HOME	Valores predeterminados

Antes de utilizar AIX `hdisk` dispositivos en cualquier entorno, incluidas las bases de datos, verifique el parámetro `queue_depth`. Este parámetro no es la profundidad de la cola HBA, sino que se relaciona con la profundidad de la cola SCSI del individuo. `hdisk device`. Dependiendo de cómo se configuren los LUN ASA r2, el valor de `queue_depth` Podría ser demasiado bajo para un buen rendimiento. Las pruebas han demostrado que el valor óptimo es 64.

HP-UX

Temas de configuración para la base de datos Oracle en HP-UX con ASA r2 ONTAP.

HP-UX es compatible con NetApp ASA r2 para alojar bases de datos Oracle, siempre que:

- La versión de ONTAP es 9.16.x o posterior.
- Utilice protocolos SAN (FC/iSCSI/NVMe). NAS no es compatible con ASA r2.
- Aplique las mejores prácticas de montaje y ajuste de E/S específicas de HP-UX.

Opciones de montaje HP-UX VxFS

Utilice las siguientes opciones de montaje para sistemas de archivos que alojan binarios de Oracle:

```
delaylog,nodatainlog
```

Utilice las siguientes opciones de montaje para sistemas de archivos que contienen archivos de datos, redo logs, archive logs y archivos de control en los que la versión de HP-UX no admite E/S simultáneas:

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

Cuando se admiten E/S simultáneas (VxFS 5.0.1 y posteriores, o con ServiceGuard Storage Management Suite), utilice estas opciones de montaje para sistemas de archivos que contengan archivos de datos, redo logs, archive logs y archivos de control:

```
delaylog,cio
```

> El parámetro `db_file_multiblock_read_count` Es especialmente crítico en entornos VxFS. Oracle recomienda que este parámetro permanezca sin definir en Oracle 10g R1 y posteriores a menos que se indique lo contrario específicamente. El valor por defecto con un tamaño de bloque de Oracle 8KB es 128. Si el valor de este parámetro se fuerza a 16 o menos, quite el `convosync=direct` Opción de montaje porque puede dañar el rendimiento de I/O secuencial. Este paso daña otros aspectos del rendimiento y solo debe tomarse si el valor de `db_file_multiblock_read_count` debe cambiarse a partir del valor predeterminado.

Linux

Temas de configuración específicos del sistema operativo Linux con ASA r2 ONTAP.

> Linux (Oracle Linux, RHEL, SUSE) es compatible con ASA r2 para bases de datos Oracle. Utilice protocolos SAN, configure correctamente las rutas múltiples y aplique las mejores prácticas de Oracle para el ajuste de ASM y E/S.

Programador de I/O.

El kernel de Linux permite un control de bajo nivel sobre la forma en que se programa la E/S para bloquear los dispositivos. Los valores por defecto en varias distribuciones de Linux varían considerablemente. Las pruebas demuestran que la fecha límite suele ofrecer los mejores resultados, pero en ocasiones NOOP ha sido ligeramente mejor. La diferencia de rendimiento es mínima, pero pruebe ambas opciones si es necesario extraer el máximo rendimiento posible de una configuración de base de datos. CFQ es el valor predeterminado en muchas configuraciones y ha demostrado tener problemas de rendimiento significativos con cargas de trabajo de bases de datos.

Consulte la documentación relevante del proveedor de Linux para obtener instrucciones sobre la configuración del programador de E/S.

Accesos múltiples

Algunos clientes se han encontrado con fallos durante la interrupción de la red porque el daemon multivía no se estaba ejecutando en su sistema. En versiones recientes de Linux, el proceso de instalación del sistema operativo y el daemon de rutas múltiples pueden dejar estos sistemas operativos vulnerables a este problema. Los paquetes están instalados correctamente, pero no están configurados para el inicio automático después de un reinicio.

Por ejemplo, el valor predeterminado para el demonio multiruta en RHEL 9.7 podría aparecer de la siguiente manera:

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

Esto se puede corregir con los siguientes comandos:

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

Profundidad de la cola

Establezca la profundidad de cola adecuada para los dispositivos SAN para evitar cuellos de botella de E/S. La profundidad de cola predeterminada en Linux a menudo se establece en 128, lo que puede generar problemas de rendimiento con las bases de datos Oracle. Establecer una profundidad de cola demasiado alta puede provocar una cola de E/S excesiva, lo que genera una mayor latencia y una reducción del rendimiento. Establecerlo demasiado bajo puede limitar la cantidad de solicitudes de E/S pendientes, lo que reduce el rendimiento general. Una profundidad de cola de 64 suele ser un buen punto de partida para las cargas de trabajo de bases de datos Oracle en ASA r2, pero es posible que sea necesario ajustarla en función de las características específicas de la carga de trabajo y las pruebas de rendimiento.

Duplicación de ASM

La duplicación de ASM puede requerir cambios en la configuración multivía de Linux para permitir que ASM reconozca un problema y cambie a un grupo de fallos alternativo. La mayoría de las configuraciones de ASM en ONTAP utilizan redundancia externa, lo que significa que la cabina externa ofrece protección de datos y ASM no refleja datos. Algunos sitios utilizan ASM con redundancia normal para proporcionar duplicación bidireccional, normalmente en diferentes sitios.

Para los sistemas ASA r2 que admiten rutas múltiples activo-activo, se deben ajustar estas configuraciones de rutas múltiples. Dado que todas las rutas están activas y tienen equilibrio de carga, no se requiere una cola indefinida. En lugar de ello, los parámetros de rutas múltiples deberían priorizar el rendimiento y la recuperación rápida. Este comportamiento es importante para la duplicación de ASM porque ASM debe recibir una falla de E/S para poder volver a intentar la E/S en un LUN alternativo. Si la E/S se pone en cola indefinidamente, ASM no puede activar una conmutación por error.

Defina los siguientes parámetros en Linux `multipath.conf` Archivo para LUN de ASM utilizados con la duplicación de ASM:

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

Estos valores crean un timeout de 120 segundos para los dispositivos ASM. El tiempo de espera se calcula como el `polling_interval` * `no_path_retry` como segundos. Puede que sea necesario ajustar el valor exacto en algunas circunstancias, pero un tiempo de espera de 120 segundos debería ser suficiente para la mayoría de los usos. Concretamente, 120 segundos deberían permitir que se produzca una toma de control o una devolución de la controladora sin que se produzca un error de I/O, lo que provocaría que el grupo de errores se desconectara.

A inferior `no_path_retry` Value puede reducir el tiempo necesario para que ASM cambie a un grupo de fallos alternativo, pero esto también aumenta el riesgo de una conmutación por error no deseada durante actividades de mantenimiento como la toma de control de un controlador. El riesgo se puede mitigar mediante una supervisión cuidadosa del estado de duplicación de ASM. Si se produce una conmutación al respaldo no deseada, los duplicados pueden volver a sincronizarse rápidamente si la resincronización se realiza con relativa rapidez. Para obtener información adicional, consulte la documentación de Oracle on ASM Fast Mirror Resync para ver la versión del software de Oracle en uso.

Opciones de montaje de Linux xfs, ext3 y ext4

* NetApp recomienda* utilizar las opciones de montaje predeterminadas. Asegúrese de que haya una alineación adecuada al crear sistemas de archivos en LUN.

ASMLib/AFD (Controlador de Filtro de ASM)

Temas de configuración específicos del sistema operativo Linux que utilizan AFD y ASMlib con ASA r2 ONTAP.

Tamaños de bloque ASMLib

ASMlib es una biblioteca de gestión de ASM opcional y utilidades asociadas. Su valor principal es la capacidad de sellar un LUN como un recurso ASM con una etiqueta legible para humanos.

Las versiones recientes de ASMLib detectan un parámetro de LUN llamado Logical Blocks per Physical Block Exponent (LBPPBE). El destino SCSI de ONTAP no notificó este valor hasta hace poco. Ahora devuelve un valor que indica que se prefiere un tamaño de bloque de 4KB KB. Esta no es una definición de tamaño de bloque, pero es una indicación para cualquier aplicación que utilice LBPPBE de que las E/S de un determinado tamaño podrían manejarse de manera más eficiente. Sin embargo, ASMLib interpreta LBPPBE como un tamaño de bloque y marca de forma persistente la cabecera ASM cuando se crea el dispositivo ASM.

Este proceso puede causar problemas con actualizaciones y migraciones de varias maneras, todo ello en función de la incapacidad de mezclar dispositivos ASMLib con diferentes tamaños de bloque en el mismo grupo de discos ASM.

Por ejemplo, las matrices más antiguas generalmente reportaron un valor LBPPBE de 0 o no reportaron este valor en absoluto. ASMLib lo interpreta como un tamaño de bloque de 512 bytes. Las cabinas más recientes se interpretarán con un tamaño de bloque de 4KB KB. No es posible mezclar dispositivos de 512 bytes y 4KB en el mismo grupo de discos ASM. Al hacerlo, se bloquearía a un usuario para que no aumente el tamaño del

grupo de discos de ASM utilizando LUN de dos matrices o aprovechando ASM como herramienta de migración. En otros casos, es posible que RMAN no permita la copia de archivos entre un grupo de discos de ASM con un tamaño de bloque de 512 bytes y un grupo de discos de ASM con un tamaño de bloque de 4KB KB.

La solución preferida es parchear ASMLib. El identificador de error de Oracle es 13999609 y el parche está presente en oracleasm-support-2,1.8-1 y superior. Este parche permite al usuario definir el parámetro `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` para `true` en la `/etc/sysconfig/oracleasm` archivo de configuración. Al hacerlo, se bloquea ASMLib para que no utilice el parámetro LBPPBE, lo que significa que los LUN de la nueva matriz ahora se reconocen como dispositivos de bloque de 512 bytes.

> La opción no cambia el tamaño de bloque en LUN que ASMLib estampó anteriormente. Por ejemplo, si un grupo de discos ASM con bloques de 512 bytes debe migrarse a un nuevo sistema de almacenamiento que notifique un bloque de 4KB KB, la opción `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Debe establecerse antes de que las nuevas LUN se estampen con ASMLib. Si los dispositivos ya han sido estampados por oracleasm, deben ser reformateados antes de ser reincorporados con un nuevo tamaño de bloque. En primer lugar, desconfigure el dispositivo con `oracleasm deletedisk`, Y, a continuación, borre los primeros 1GB del dispositivo con `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Por último, si el dispositivo se ha particionado previamente, utilice `kpartx` Comando para eliminar particiones obsoletas o simplemente reiniciar el sistema operativo.

Si no se puede aplicar un parche a ASMLib, se puede eliminar ASMLib de la configuración. Este cambio es disruptivo y requiere el desestampado de discos de ASM y asegurarse de que el `asm_diskstring` el parámetro se ha definido correctamente. Sin embargo, este cambio no requiere la migración de datos.

Tamaños de bloque de unidad de filtro de ASM (AFD)

AFD es una biblioteca de gestión de ASM opcional que se está convirtiendo en el reemplazo de ASMLib. Desde el punto de vista del almacenamiento, es muy similar a ASMLib, pero incluye características adicionales como la capacidad de bloquear E/S no Oracle para reducir las posibilidades de errores de usuario o aplicación que podrían dañar los datos.

Tamaños de bloques de dispositivos

Al igual que ASMLib, AFD también lee el parámetro LUN Bloques lógicos por Exponente de bloque físico (LBPPBE) y utiliza de forma predeterminada el tamaño del bloque físico, no el tamaño del bloque lógico.

Esto podría crear un problema si se agrega AFD a una configuración existente donde los dispositivos ASM ya están formateados como dispositivos de bloque de 512 bytes. El controlador AFD reconocería el LUN como un dispositivo 4K y la discrepancia entre la etiqueta ASM y el dispositivo físico impediría el acceso. Del mismo modo, las migraciones se verían afectadas porque no es posible mezclar dispositivos de 512 bytes y 4KB en el mismo grupo de discos de ASM. Al hacerlo, se bloquearía a un usuario para que no aumente el tamaño del grupo de discos de ASM utilizando LUN de dos matrices o aprovechando ASM como herramienta de migración. En otros casos, es posible que RMAN no permita la copia de archivos entre un grupo de discos de ASM con un tamaño de bloque de 512 bytes y un grupo de discos de ASM con un tamaño de bloque de 4KB KB.

La solución es simple: AFD incluye un parámetro para controlar si utiliza los tamaños de bloque lógicos o físicos. Este es un parámetro global que afecta a todos los dispositivos del sistema. Para forzar a AFD a utilizar el tamaño de bloque lógico, establezca `options oracleafd oracleafd_use_logical_block_size=1` en la `/etc/modprobe.d/oracleafd.conf` archivo.

Tamaños de transferencia multivía

Los cambios recientes del kernel de linux aplican las restricciones de tamaño de I/O enviadas a dispositivos multivía y el AFD no cumple con estas restricciones. A continuación, se rechazan las I/O, lo que hace que la ruta de LUN se desconecte. El resultado es una incapacidad para instalar Oracle Grid, configurar ASM o crear una base de datos.

La solución es especificar manualmente la longitud máxima de transferencia del archivo multipath.conf para las LUN de ONTAP:

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

Incluso si no existe ningún problema en la actualidad, este parámetro debe configurarse si se utiliza AFD para garantizar que una actualización de linux futura no cause problemas de forma inesperada.

Microsoft Windows

Temas de configuración para la base de datos Oracle en Microsoft Windows con ASA r2 ONTAP.

SAN

Para una eficiencia de compresión óptima, asegúrese de que el sistema de archivos NTFS utilice una unidad de asignación de 8K o más. El uso de una unidad de asignación de 4K, que suele ser la predeterminada, afecta negativamente a la eficiencia de la compresión.

Solaris

Temas de configuración específicos del sistema operativo Solaris con ASA r2 ONTAP.

Opciones de montaje UFS de Solaris

NetApp recomienda usar la opción de montaje de registro para conservar la integridad de los datos en caso de bloqueo del host Solaris o interrupción de la conectividad de FC. La opción de montaje logging también conserva la facilidad de uso de los backups de Snapshot.

ZFS de Solaris

Solaris ZFS debe instalarse y configurarse cuidadosamente para ofrecer un rendimiento óptimo.

mvector

Solaris 11 incluyó un cambio en la forma en que procesa grandes operaciones de E/S, lo que puede dar lugar

a graves problemas de rendimiento en las matrices de almacenamiento SAN. El problema está documentado en el informe de errores de seguimiento de NetApp 630173, que indica que Solaris 11 ZFS Performance Regression.

Esto no es un bug de ONTAP. Se trata de un defecto de Solaris cuyo seguimiento se realiza bajo los defectos de Solaris 7199305 y 7082975.

Puede consultar a los Servicios de Soporte Oracle para averiguar si la versión de Solaris 11 está afectada o puede probar la solución provisional cambiando `zfs_mvector_max_size` a un valor menor.

Para ello, ejecute el siguiente comando como root:

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

Si surge algún problema inesperado de este cambio, se puede revertir fácilmente ejecutando el siguiente comando como root:

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

Kernel

El rendimiento fiable de ZFS requiere un kernel de Solaris parcheado contra problemas de alineación de LUN. La corrección se introdujo con el parche 147440-19 en Solaris 10 y con SRU 10,5 para Solaris 11. Utilice sólo Solaris 10 y versiones posteriores con ZFS.

Configuración de LUN

Para configurar una LUN, complete los siguientes pasos:

1. Cree una LUN del tipo `solaris`.
2. Instale el kit de utilidades de host (HUK) adecuado especificado por el "Herramienta de matriz de interoperabilidad de NetApp (IMT)".
3. Siga las instrucciones del HUK exactamente como se describe. Los pasos básicos se describen a continuación, pero consulte la "documentación más reciente" para el procedimiento adecuado.
   a. Ejecute el `host_config` utilidad para actualizar el `sd.conf/sdd.conf` archivo. Al hacerlo, las unidades SCSI pueden detectar correctamente LUN de ONTAP.
   b. Siga las instrucciones proporcionadas por el `host_config` Utilidad para habilitar la entrada/salida multivía (MPIO).
   c. Reiniciar. Este paso es necesario para que cualquier cambio se reconozca en el sistema.
4. Cree particiones en las LUN y compruebe que están correctamente alineadas. Consulte el Apéndice B: Verificación de alineación de WAFL para obtener instrucciones sobre cómo probar y confirmar la alineación directamente.

zpools

Sólo se debe crear un zpool después de los pasos del "Configuración de LUN" se realizan. Si el procedimiento no se realiza correctamente, puede provocar una degradación grave del rendimiento debido a la alineación de E/S. Para un rendimiento óptimo en ONTAP es necesario alinear el I/O con un límite de 4K GbE en una

unidad. Los sistemas de archivos creados en un zpool utilizan un tamaño de bloque efectivo que se controla mediante un parámetro denominado `ashift`, que se puede ver ejecutando el comando `zdb -C`.

Valor de `ashift` el valor por defecto es 9, que significa 2^9, o 512 bytes. Para un rendimiento óptimo, el `ashift` El valor debe ser 12 (2^12=4K). Este valor se define en el momento en que se crea zpool y no se puede cambiar, lo que significa que los datos en zpools con `ashift` los datos que no sean 12 se deben migrar copiando a un zpool recién creado.

Después de crear un zpool, verifique el valor de `ashift` antes de continuar. Si el valor no es 12, las LUN no se detectaron correctamente. Destruya zpool, verifique que todos los pasos mostrados en la documentación de utilidades de host relevantes se hayan realizado correctamente y vuelva a crear zpool.

Zpools y LDOMs de Solaris

Los LDOMs de Solaris crean un requisito adicional para asegurarse de que la alineación de E/S es correcta. Aunque un LUN se puede detectar correctamente como dispositivo 4K, un dispositivo virtual vdsk en un LDOM no hereda la configuración del dominio de E/S. El vdsk basado en esa LUN vuelve a tener de forma predeterminada un bloque de 512 bytes.

Se necesita un archivo de configuración adicional. En primer lugar, se deben aplicar parches a los LDOM individuales para el bug de Oracle 15824910 para activar las opciones de configuración adicionales. Este parche se ha portado a todas las versiones utilizadas actualmente de Solaris. Una vez que se aplica el parche a LDOM, está listo para la configuración de las nuevas LUN correctamente alineadas de la siguiente manera:

1. Identifique los LUN o LUN que se van a utilizar en el nuevo zpool. En este ejemplo, es el dispositivo c2d1.

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

2. Recuperar la instancia vdc de los dispositivos que se van a utilizar para una agrupación ZFS:

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

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

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

Esto significa que a la instancia de dispositivo 1 se le asigna un tamaño de bloque de 4096.

Como ejemplo adicional, supongamos que las instancias de vdsk 1 a 6 deben configurarse para un tamaño de bloque de 4K KB y. `/etc/path_to_inst` se lee de la siguiente manera:

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

4. La final `vdc.conf` el archivo debe contener lo siguiente:

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

El LDOM debe reiniciarse después de configurar vdc.conf y crear vdsk. Este paso no se puede evitar. El cambio de tamaño del bloque solo se aplica después de un reinicio. Continúe con la configuración de zpool y asegúrese de que el ashift está correctamente ajustado en 12 como se ha descrito anteriormente.

Registro de Intención de ZFS (ZIL)

Por lo general, no hay razón para localizar el registro de intención ZFS (ZIL) en un dispositivo diferente. El registro puede compartir espacio con el pool principal. El uso principal de un ZIL separado es cuando se utilizan unidades físicas que carecen de las funciones de almacenamiento en caché de escritura en cabinas de almacenamiento modernas.

sesgo logarítmico

Ajuste la `logbias` Parámetro en sistemas de archivos ZFS que alojan datos de Oracle.

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

Usar este parámetro reduce los niveles generales de escritura. En los valores predeterminados, los datos escritos se confirman primero en el ZIL y, a continuación, en el pool de almacenamiento principal. Este enfoque es adecuado para una configuración que utiliza una configuración de unidad simple, que incluye un dispositivo ZIL basado en SSD y medios giratorios para el pool de almacenamiento principal. Esto se debe a que permite un commit en una sola transacción de I/O en el medio de menor latencia disponible.

Cuando se utiliza una cabina de almacenamiento moderna que incluye su propia funcionalidad de almacenamiento en caché, este método no suele ser necesario. En raras ocasiones, es posible que sea conveniente comprometer una escritura con una sola transacción en el registro, como una carga de trabajo que consta de escrituras aleatorias altamente concentradas y sensibles a la latencia. Existen consecuencias en la amplificación de escritura, ya que los datos registrados se escriben finalmente en el pool de almacenamiento principal, lo que provoca el doble de la actividad de escritura.

E/S directa

Muchas aplicaciones, incluidos los productos de Oracle, pueden omitir la caché de buffers del host activando la E/S directa Esta estrategia no funciona como se esperaba con los sistemas de archivos ZFS. Aunque se omite la caché de buffers del host, ZFS continúa almacenando los datos en caché. Esta acción puede provocar resultados engañosos cuando se usan herramientas como fio o sio para realizar pruebas de rendimiento, ya que es difícil predecir si I/O está llegando al sistema de almacenamiento o si se está almacenando en caché localmente dentro del sistema operativo. Esta acción también hace que sea muy difícil utilizar estas pruebas sintéticas para comparar el rendimiento de ZFS con otros sistemas de archivos. Como cuestión práctica, hay poca o ninguna diferencia en el rendimiento del sistema de archivos con las cargas de trabajo de los usuarios reales.

Varios zpools

Las copias de seguridad basadas en instantáneas, las restauraciones, los clones y el archivado de datos basados en ZFS se deben realizar en el nivel de zpool y, por lo general, requieren varios zpools. Un zpool es análogo a un grupo de discos LVM y debe configurarse usando las mismas reglas. Por ejemplo, es probable que una base de datos se disponga mejor con los archivos de datos en los que reside `zpool1` y los registros de archivo, los archivos de control y los registros de recuperación en los que residen `zpool2`. Este enfoque permite realizar un backup dinámico estándar en el que la base de datos se coloca en modo de backup dinámico, seguido de una copia Snapshot de `zpool1`. A continuación, la base de datos se elimina del modo

de backup dinámico, se fuerza el archivo de registro y una copia de Snapshot de `zpool2` se ha creado. Una operación de restauración requiere el desmontaje de los sistemas de archivos zfs y desconectar zpool íntegramente, a continuación de una operación de restauración de SnapRestore. El zpool se puede poner en línea de nuevo y la base de datos se recupera.

filesystemio_options

Parámetro de Oracle `filesystemio_options` Funciona de forma diferente con ZFS. Si `setall` o. `directio` Se utiliza, las operaciones de escritura son síncronas y omiten la caché de buffers del sistema operativo, pero ZFS almacena en búfer las lecturas. Esta acción causa dificultades en el análisis de rendimiento porque a veces la caché ZFS intercepta y suministra servicio a las E/S, lo que hace que la latencia de almacenamiento y el total de E/S sean menores de lo que podría parecer.

Configuración de red en sistemas AFF/ FAS

Interfaces lógicas

Las bases de datos de Oracle necesitan acceder al almacenamiento. Las interfaces lógicas (LIF) son las tuberías de red que conecta una máquina virtual de almacenamiento (SVM) a la red y, por tanto, a la base de datos. Es necesario diseñar un LIF adecuado para garantizar que exista un ancho de banda suficiente para cada carga de trabajo de la base de datos. La conmutación por error no conlleva la pérdida de los servicios de almacenamiento.

En esta sección se ofrece una descripción general de los principios clave del diseño de LIF. Para obtener documentación más completa, consulte "Documentación de gestión de red de ONTAP". Al igual que otros aspectos de la arquitectura de bases de datos, las mejores opciones para las máquinas virtuales de almacenamiento (SVM, conocidas como Vserver en la CLI) y el diseño de la interfaz lógica (LIF) dependen en gran medida de los requisitos de escalado y las necesidades empresariales.

Tenga en cuenta los siguientes temas principales al crear una estrategia de LIF:

- **Rendimiento.** ¿Es suficiente el ancho de banda de la red?
- **Resiliencia.** ¿Hay algún punto de falla en el diseño?
- **Capacidad de gestión.** ¿Se puede escalar la red de forma no disruptiva?

Estos temas se aplican a la solución completa, desde el host, los switches y el sistema de almacenamiento.

Tipos de LIF

Hay varios tipos de LIF. "Documentación de ONTAP sobre tipos de LIF" Puede proporcionar información más completa sobre este tema, pero desde una perspectiva funcional, los LIF se pueden dividir en los siguientes grupos:

- **LIF de administración de clúster y nodos.** LIF utilizadas para administrar el clúster de almacenamiento.
- **LIF de administración de SVM.** Interfaces que permiten el acceso a una SVM a través de la API REST o ONTAPI (también conocida como ZAPI) para funciones como la creación de instantáneas o el redimensionamiento de volúmenes. Productos como SnapManager para Oracle (SMO) deben tener acceso a una LIF de gestión de SVM.
- **LIF de datos.** Interfaces para FC, iSCSI, NVMe/FC, NVMe/TCP, NFS, o datos SMB/CIFS.

También puede utilizarse una LIF de datos que se utiliza para el tráfico NFS al cambiar la política de firewall de `data` para `mgmt` O cualquier otra política que permita HTTP, HTTPS o SSH. Este cambio puede simplificar la configuración de red ya que evita la configuración de cada host para obtener acceso a tanto a la LIF de datos de NFS como a una LIF de gestión separada. No se puede configurar una interfaz para iSCSI y para el tráfico de gestión, a pesar de que ambos usen un protocolo IP. En los entornos iSCSI, se requiere una LIF de gestión separada.

Diseño de LIF SAN

El diseño de LIF en un entorno SAN es relativamente sencillo por una de las razones: La multivía. Todas las implementaciones de SAN modernas permiten a un cliente acceder a los datos a través de múltiples rutas de red independientes y seleccionar la mejor ruta o las mejores rutas para acceder. Como resultado, el rendimiento con respecto al diseño de las LIF es más sencillo de abordar porque los clientes SAN equilibran automáticamente la carga de I/O en las mejores rutas disponibles.

Si una ruta deja de estar disponible, el cliente selecciona automáticamente una ruta diferente. La simplicidad resultante del diseño hace que los LIF SAN sean generalmente más gestionables. Esto no significa que un entorno SAN siempre se gestione con mayor facilidad, ya que existen otros muchos aspectos del almacenamiento SAN que son mucho más complicados que NFS. Simplemente significa que el diseño de LIF SAN es más sencillo.

Rendimiento

El aspecto más importante con respecto al rendimiento de LIF en un entorno SAN es el ancho de banda. Por ejemplo, un clúster ONTAP AFF de dos nodos con dos puertos FC de 16GB Gb por nodo permite hasta 32GB Gbps de ancho de banda hacia/desde cada nodo.

Resiliencia

Los LIF DE SAN no conmutan al nodo de respaldo en un sistema de almacenamiento AFF. Si falla un LIF de SAN debido a la recuperación tras fallos de la controladora, el software multivía del cliente detecta la pérdida de una ruta y redirige las I/O a otro LIF. Con los sistemas de almacenamiento de ASA, los LIF conmutarán por error tras un breve retraso, pero esto no interrumpe las I/O porque ya hay rutas activas en la otra controladora. El proceso de conmutación por error tiene lugar para restaurar el acceso de host en todos los puertos definidos.

Gran capacidad de administración

La migración de LIF es una tarea mucho más común en un entorno NFS porque la migración de LIF suele asociarse con la reubicación de volúmenes en el clúster. No es necesario migrar un LIF en un entorno SAN cuando se reubican volúmenes en el par de alta disponibilidad. Esto se debe a que, una vez finalizado el movimiento de volúmenes, ONTAP envía una notificación a la SAN sobre un cambio en las rutas y los clientes SAN vuelven a optimizarse automáticamente. La migración de LIF con SAN está asociada principalmente a los grandes cambios de hardware físico. Por ejemplo, si es necesaria una actualización sin interrupciones de las controladoras, se migra un LIF SAN al nuevo hardware. Si se encuentra que un puerto FC está defectuoso, puede migrarse un LIF a un puerto no utilizado.

Recomendaciones de diseño

NetApp hace las siguientes recomendaciones:

- No cree más rutas de las necesarias. Un número excesivo de rutas complica la gestión general y puede provocar problemas en la conmutación al nodo de respaldo de rutas en algunos hosts. Además, algunos

hosts tienen limitaciones inesperadas de la ruta para configuraciones como el arranque SAN.

- Muy pocas configuraciones deberían requerir más de cuatro rutas a una LUN. El valor de tener más de dos nodos de rutas de publicidad para los LUN es limitado porque no se puede acceder al agregado que aloja una LUN si se produce un error en el nodo propietario de la LUN y su partner de alta disponibilidad. La creación de rutas en nodos que no sean el par de alta disponibilidad primario no es útil en esta situación.
- Aunque puede gestionar el número de rutas visibles de LUN si selecciona qué puertos se incluyen en las zonas FC, suele ser más fácil incluir todos los puntos de destino potenciales en la zona FC y controlar la visibilidad de la LUN a nivel de ONTAP.
- En ONTAP 8,3 y versiones posteriores, la función de asignación selectiva de LUN (SLM) es la opción predeterminada. Con SLM, cualquier nuevo LUN se anuncia automáticamente desde el nodo propietario del agregado subyacente y el partner de alta disponibilidad del nodo. Esta disposición evita la necesidad de crear conjuntos de puertos o configurar la división en zonas para limitar la accesibilidad del puerto. Cada LUN está disponible en el número mínimo de nodos necesarios, tanto para un rendimiento óptimo como para una resiliencia.
  *En el caso de que una LUN deba migrarse fuera de los dos controladores, los nodos adicionales se pueden agregar con el `lun mapping add-reporting-nodes` Comando para que las LUN se anuncien en los nodos nuevos. Al hacerlo, se crean rutas de SAN adicionales a las LUN para la migración de la LUN. Sin embargo, el host debe realizar una operación de detección para utilizar las rutas nuevas.
- No se preocupe demasiado por el tráfico indirecto. Es mejor evitar el tráfico indirecto en un entorno con un gran volumen de I/O para el que cada microsegundo de latencia es crucial, pero el efecto de rendimiento visible es insignificante para las cargas de trabajo típicas.

Diseño de LIF NFS

A diferencia de los protocolos SAN, NFS tiene una capacidad limitada de definir varias rutas para los datos. Las extensiones paralelas de NFS (pNFS) instaladas en NFSv4 solucionan esta limitación, pero, como las velocidades de ethernet han alcanzado los 100GB GbE y más allá, rara vez hay valor en añadir rutas adicionales.

Rendimiento y resiliencia

Aunque medir el rendimiento de LIF de SAN se trata, principalmente, de calcular el ancho de banda total de todas las rutas principales, determinar el rendimiento de LIF NFS requiere observar con más detenimiento la configuración de red exacta. Por ejemplo, se pueden configurar dos puertos 10Gb GbE como puertos físicos sin configurar o como grupo de interfaces del protocolo de control de agregación de enlaces (LACP). Si se configuran como un grupo de interfaces, hay varias políticas de equilibrio de carga disponibles que funcionan de forma diferente en función de si el tráfico se conmuta o se enruta. Por último, Oracle Direct NFS (dNFS) ofrece configuraciones de equilibrio de carga que no existen en ningún cliente NFS del sistema operativo en este momento.

A diferencia de los protocolos SAN, los sistemas de archivos NFS requieren resiliencia en la capa de protocolo. Por ejemplo, un LUN siempre está configurado con multivía habilitado, lo que significa que hay varios canales redundantes disponibles para el sistema de almacenamiento, cada uno de los cuales utiliza el protocolo FC. Un sistema de archivos NFS, por otro lado, depende de la disponibilidad de un único canal TCP/IP que solo se puede proteger en la capa física. Esta disposición es el motivo por el cual existen opciones como la conmutación por error de puerto y la agregación de puertos LACP.

En un entorno NFS, se proporciona rendimiento y flexibilidad en la capa de protocolo de red. Como resultado, ambos temas están entrelazados y deben discutirse juntos.

Enlace las LIF a grupos de puertos

Para enlazar una LIF a un grupo de puertos, asocie la dirección IP de LIF con un grupo de puertos físicos. El principal método para añadir puertos físicos juntos es LACP. La funcionalidad de tolerancia a fallos de LACP es bastante sencilla; cada puerto de un grupo de LACP se supervisa y se elimina del grupo de puertos en caso de que se produzca un funcionamiento incorrecto. No obstante, existen muchos conceptos erróneos sobre cómo funciona LACP con respecto al rendimiento:

- LACP no requiere que la configuración del switch coincida con el extremo. Por ejemplo, ONTAP puede configurarse con balanceo de carga basado en IP, mientras que un switch puede utilizar balanceo de carga basado en MAC.
- Cada punto final que utiliza una conexión LACP puede elegir de forma independiente el puerto de transmisión de paquetes, pero no puede elegir el puerto utilizado para la recepción. Esto significa que el tráfico de ONTAP a un destino en particular está vinculado a un puerto en particular, y el tráfico de retorno podría llegar a una interfaz diferente. Sin embargo, esto no causa problemas.
- LACP no distribuye el tráfico de manera uniforme en todo momento. En un entorno de gran tamaño con muchos clientes NFS, el resultado suele utilizarse incluso en todos los puertos de una agregación de LACP. Sin embargo, cualquier sistema de archivos NFS en el entorno está limitado al ancho de banda de un solo puerto, no a toda la agregación.
- Si bien las políticas LACP de robin-robin están disponibles en ONTAP, estas políticas no abordan la conexión desde un switch a un host. Por ejemplo, una configuración con un tronco LACP de cuatro puertos en un host y un tronco LACP de cuatro puertos en ONTAP solo puede leer un sistema de archivos utilizando un único puerto. Aunque ONTAP puede transmitir datos a través de los cuatro puertos, actualmente no hay tecnologías de switches disponibles que se envíen del switch al host a través de los cuatro puertos. Solo se utiliza uno.

El enfoque más común en entornos de mayor tamaño que consisten en muchos hosts de base de datos es crear un agregado LACP de un número adecuado de interfaces 10Gb (o más rápidas) mediante el equilibrio de carga de IP. Este enfoque permite a ONTAP ofrecer un uso uniforme de todos los puertos, siempre y cuando existan suficientes clientes. El equilibrio de carga se desglosa cuando hay menos clientes en la configuración porque la conexión troncal LACP no redistribuye la carga de forma dinámica.

Cuando se establece una conexión, el tráfico en una dirección determinada se coloca en un solo puerto. Por ejemplo, una base de datos que realiza una exploración de tabla completa en un sistema de archivos NFS conectado a través de un tronco LACP de cuatro puertos lee los datos aunque solo una tarjeta de interfaz de red (NIC). Si sólo hay tres servidores de base de datos en un entorno de este tipo, es posible que los tres estén leyendo desde el mismo puerto, mientras que los otros tres puertos estén inactivos.

Enlazar LIF a puertos físicos

La vinculación de una LIF a un puerto físico provoca un control más granular sobre la configuración de red, ya que una dirección IP determinada en un sistema ONTAP solo está asociada con un puerto de red a la vez. A continuación, la resiliencia se lleva a cabo mediante la configuración de grupos de conmutación al respaldo y las políticas de conmutación por error.

Políticas de conmutación por error y grupos de conmutación por error

El comportamiento de las LIF durante la interrupción de la red está controlado por las políticas de conmutación por error y los grupos de recuperación tras fallos. Las opciones de configuración han cambiado con las distintas versiones de ONTAP. Consulte la "Documentación de gestión de redes de ONTAP para políticas y grupos de conmutación por error" Para obtener detalles específicos de la versión de ONTAP que se va a poner en marcha.

ONTAP 8,3 y superiores permiten la gestión de recuperación tras fallos de LIF en función de dominios de retransmisión. Por lo tanto, un administrador puede definir todos los puertos que tienen acceso a una subred determinada y permitir que ONTAP seleccione una LIF de conmutación al nodo de respaldo adecuada. Algunos clientes pueden utilizar este enfoque, pero tiene limitaciones en un entorno de red de almacenamiento de alta velocidad debido a la falta de previsibilidad. Por ejemplo, un entorno puede incluir ambos puertos 1GB para acceso rutinario al sistema de archivos y puertos 10Gb para las operaciones de I/O del archivo de datos Si ambos tipos de puertos existen en el mismo dominio de retransmisión, la conmutación por error de LIF puede provocar que se muevan las operaciones de I/O del archivo de datos de un puerto 10Gb a un puerto 1GB.

En resumen, tenga en cuenta las siguientes prácticas:

1. Configure un grupo de failover como definido por el usuario.
2. Rellenar el grupo de recuperación tras fallos con puertos en el controlador asociado de recuperación tras fallos de almacenamiento (SFO) de modo que los LIF sigan a los agregados durante una conmutación al nodo de respaldo de almacenamiento. Esto evita la creación de tráfico indirecto.
3. Utilice puertos de conmutación por error con las características de rendimiento correspondientes a la LIF original. Por ejemplo, un LIF en un único puerto físico 10Gb debería incluir un grupo de conmutación por error con un único puerto 10Gb. Un LIF LACP de cuatro puertos debe conmutar por error a otro LIF LACP de cuatro puertos. Estos puertos serían un subconjunto de los puertos definidos en el dominio de retransmisión.
4. Establezca la política de recuperación tras fallos únicamente en SFO-partner. Al hacerlo, se asegura de que el LIF siga al agregado durante la recuperación tras fallos.

Reversión automática

Ajuste la `auto-revert` parámetro como desee. La mayoría de los clientes prefieren establecer este parámetro en `true` Para que la LIF vuelva a su puerto de inicio. Sin embargo, en algunos casos, los clientes han establecido esto en 'false' para que se pueda investigar una conmutación por error inesperada antes de devolver una LIF a su puerto de origen.

Proporción de LIF a volumen

Un concepto erróneo común es que debe haber una relación de 1:1 GbE entre los volúmenes y los LIF de NFS. Aunque esta configuración es necesaria para mover un volumen a cualquier punto de un clúster mientras no se crea tráfico de interconexión adicional, no es categóricamente un requisito. Hay que tener en cuenta el tráfico entre clústeres, pero la mera presencia del tráfico entre clústeres no crea problemas. Muchas de las pruebas de rendimiento publicadas creadas para ONTAP incluyen I/O predominantemente indirectas

Por ejemplo, un proyecto de base de datos que contiene una cantidad relativamente pequeña de bases de datos críticas para el rendimiento que solo requerían un total de 40 volúmenes podría justificar un volumen de 1:1 GB para la estrategia LIF, una disposición que requeriría 40 direcciones IP. Posteriormente, cualquier volumen se podría mover a cualquier parte del clúster junto con la LIF asociada; el tráfico siempre sería directo, minimizando todas las fuentes de latencia incluso a niveles de microsegundos.

Como ejemplo por contador, un entorno alojado de gran tamaño se podría gestionar más fácilmente con una relación de 1:1:1 entre clientes y las LIF. Con el tiempo, es posible que se deba migrar un volumen a un nodo diferente, lo cual provocaría cierto tráfico indirecto. Sin embargo, el efecto de rendimiento debe ser indetectable a menos que los puertos de red en el conmutador de interconexión estén saturados. Si hay algún problema, se puede establecer un nuevo LIF en nodos adicionales y el host puede actualizarse en la siguiente ventana de mantenimiento para eliminar el tráfico indirecto de la configuración.

Configuración TCP/IP y ethernet

Muchos clientes de Oracle en ONTAP utilizan ethernet, el protocolo de red de NFS, iSCSI, NVMe/TCP y, especialmente, el cloud.

Configuración del sistema operativo host

La mayoría de la documentación del proveedor de aplicaciones incluye configuraciones TCP y ethernet específicas para garantizar que la aplicación funcione de manera óptima. Estas mismas configuraciones suelen ser suficientes para ofrecer también un rendimiento óptimo del almacenamiento basado en IP.

Control de flujo Ethernet

Esta tecnología permite a un cliente solicitar que un remitente detenga temporalmente la transmisión de datos. Esto suele hacerse porque el receptor no puede procesar los datos entrantes con la suficiente rapidez. Al mismo tiempo, solicitar que un remitente cesara la transmisión era menos perjudicial que tener un receptor descarte de paquetes porque los buffers estaban llenos. Este ya no es el caso con las pilas TCP utilizadas en los sistemas operativos actualmente. De hecho, el control de flujo causa más problemas de los que resuelve.

Los problemas de rendimiento causados por el control de flujo de Ethernet han aumentado en los últimos años. Esto se debe a que el control de flujo Ethernet funciona en la capa Physical. Si una configuración de red permite que un sistema operativo del host envíe una solicitud de control de flujo de Ethernet a un sistema de almacenamiento, el resultado es una pausa en las operaciones de I/O de todos los clientes conectados. Debido a que una única controladora de almacenamiento atiende cada vez más a un número de clientes, la probabilidad de que uno o varios de estos clientes envíen solicitudes de control de flujo aumenta. El problema se ha observado con frecuencia en las instalaciones de los clientes con una amplia virtualización del SO.

Una NIC de un sistema NetApp no debe recibir solicitudes de control de flujo. El método utilizado para lograr este resultado varía según el fabricante del conmutador de red. En la mayoría de los casos, el control de flujo en un conmutador Ethernet se puede establecer en `receive desired` o. `receive on`, lo que significa que una solicitud de control de flujo no se reenvía al controlador de almacenamiento. En otros casos, la conexión de red en la controladora de almacenamiento puede no permitir la deshabilitación de control de flujo. En estos casos, los clientes deben configurarse para que nunca envíen solicitudes de control de flujo, ya sea cambiando a la configuración de NIC en el propio servidor host o a los puertos de switch a los que está conectado el servidor host.

NetApp recomienda asegurarse de que los controladores de almacenamiento NetApp no reciban paquetes de control de flujo Ethernet. Por lo general, esto puede realizarse mediante la configuración de los puertos del switch a los que está conectada la controladora, pero algunas limitaciones en el hardware del switch pueden requerir cambios en el lado del cliente.

Tamaños de MTU

Se ha demostrado que el uso de tramas gigantes ofrece alguna mejora del rendimiento en las redes 1GB al reducir la sobrecarga de la CPU y de la red, pero el beneficio no suele ser significativo.

NetApp recomienda implementar marcos jumbo cuando sea posible, tanto para obtener beneficios potenciales de rendimiento como para preparar la solución para el futuro.

El uso de tramas gigantes en una red 10Gb es casi obligatorio. Esto se debe a que la mayoría de las implementaciones de 10Gb alcanzan un límite de paquetes por segundo sin tramas gigantes antes de alcanzar la marca de 10Gb. El uso de tramas gigantes mejora la eficiencia del procesamiento TCP/IP porque permite que el sistema operativo, el servidor, las NIC y el sistema de almacenamiento procesen menos

paquetes, pero más grandes. La mejora del rendimiento varía de NIC a NIC, pero es significativa.

En las implementaciones de tramas gigantes, existe la creencia común, aunque incorrecta, de que todos los dispositivos conectados deben admitir tramas gigantes y que el tamaño de MTU debe coincidir de extremo a extremo En su lugar, los dos extremos de red negocian el tamaño de trama más alto mutuamente aceptable al establecer una conexión. En un entorno típico, un switch de red se establece con un tamaño de MTU de 9216, la controladora NetApp se establece en 9000 y los clientes se configuran con una combinación de 9000 y 1514. Los clientes que admiten un MTU de 9000 pueden utilizar tramas gigantes, y los clientes que solo puedan admitir 1514 pueden negociar un valor inferior.

Los problemas con esta disposición son raros en un entorno completamente conmutado. Sin embargo, tenga cuidado en un entorno enrutado que ningún enrutador intermedio se vea forzado a fragmentar tramas gigantes.

NetApp recomienda configurar lo siguiente:

- Las tramas gigantes son deseables, pero no se requieren con Ethernet de 1GB Gb (GbE).
- Se requieren tramas gigantes para lograr el máximo rendimiento, con 10GbE y más rápido.

Parámetros de TCP

A menudo hay tres ajustes mal configurados: Marcas de tiempo TCP, reconocimiento selectivo (SACK) y escalado de ventana TCP. Muchos documentos desactualizados en Internet recomiendan deshabilitar uno o varios de estos parámetros para mejorar el rendimiento. Había algo de mérito en esta recomendación hace muchos años, cuando las capacidades de la CPU eran mucho menores y había un beneficio en reducir la sobrecarga en el procesamiento TCP siempre que fuera posible.

Sin embargo, con los sistemas operativos modernos, deshabilitar cualquiera de estas características de TCP generalmente no resulta en ningún beneficio detectable, a la vez que también puede dañar el rendimiento. Los daños en el rendimiento son especialmente probables en entornos de red virtualizados, ya que estas características son necesarias para gestionar eficazmente la pérdida de paquetes y los cambios en la calidad de la red.

NetApp recomienda habilitar las marcas de tiempo TCP, EL SACK y el escalado de la ventana TCP en el host, y los tres parámetros deben estar activados por defecto en cualquier sistema operativo actual.

Configuración de SAN FC

La configuración de SAN FC para bases de datos de Oracle consiste principalmente en seguir prácticas recomendadas diarias de SAN.

Esto incluye medidas de planificación típicas, como asegurar que exista suficiente ancho de banda en la SAN entre el host y el sistema de almacenamiento, comprobar que existan todas las rutas de SAN entre todos los dispositivos requeridos, mediante la configuración de puertos FC requerida por el proveedor de switches FC, para evitar la contención de ISL, y con una supervisión adecuada del tejido SAN.

División en zonas

Una zona de FC nunca debe contener más de un iniciador. Tal arreglo puede parecer funcionar inicialmente, pero la comunicación entre iniciadores finalmente interfiere con el rendimiento y la estabilidad.

Las zonas multidestino se consideran generalmente seguras, aunque en raras ocasiones el comportamiento

de los puertos de destino FC de diferentes proveedores ha causado problemas. Por ejemplo, evite incluir los puertos de destino de una cabina de almacenamiento NetApp y otra que no sea de NetApp en la misma zona. Además, es aún más probable que la ubicación de un sistema de almacenamiento NetApp y un dispositivo de cinta en la misma zona cause problemas.

Red de conexión directa

A veces, los administradores de almacenamiento prefieren simplificar sus infraestructuras eliminando los switches de red de la configuración. Esto puede ser soportado en algunos escenarios.

ISCSI y NVMe/TCP

Un host que utilice iSCSI o NVMe/TCP se puede conectar directamente a un sistema de almacenamiento y funcionar normalmente. El motivo son las rutas. Las conexiones directas a dos controladoras de almacenamiento diferentes dan como resultado dos rutas independientes para el flujo de datos. La pérdida de una ruta, un puerto o una controladora no impide que se utilice la otra ruta.

NFS

Se puede utilizar el almacenamiento NFS conectado directamente, pero con una limitación considerable: El fallo no funcionará si no se realiza una ejecución significativa de secuencias de comandos, que sería responsabilidad del cliente.

El motivo por el que la recuperación tras fallos sin interrupciones se complica gracias al almacenamiento NFS de conexión directa es el enrutamiento que se produce en el sistema operativo local. Por ejemplo, supongamos que un host tiene una dirección IP de 192.168.1.1/24 y está directamente conectado a una controladora ONTAP con la dirección IP 192.168.1.50/24. Durante la conmutación al nodo de respaldo, esa dirección 192.168.1.50 puede conmutar al nodo de respaldo a la otra controladora y estará disponible para el host, pero ¿cómo detecta el host su presencia? La dirección 192.168.1.1 original todavía existe en la NIC host que ya no se conecta a un sistema operativo. El tráfico destinado a 192.168.1.50 seguiría enviándose a un puerto de red inoperable.

La segunda NIC del SO podría configurarse como 19 2.168.1.2 y sería capaz de comunicarse con la dirección fallida en 192.168.1.50, pero las tablas de enrutamiento locales tendrían un valor predeterminado de usar una dirección **y solo una** para comunicarse con la subred 192.168.1.0/24. Un administrador de sistema podría crear un marco de scripting que detectara una conexión de red fallida y alterara las tablas de enrutamiento locales o activara o desactivara las interfaces. El procedimiento exacto dependerá del sistema operativo en uso.

En la práctica, los clientes de NetApp disponen de NFS conectado directamente, pero normalmente solo para cargas de trabajo en las que se pueden pausar I/O durante las recuperaciones tras fallos. Cuando se utilizan montajes duros, no debe haber ningún error de E/S durante dichas pausas. El I/O se debe bloquear hasta que los servicios se restauren, ya sea mediante una conmutación de retorno tras recuperación o intervención manual para mover las direcciones IP entre las NIC del host.

Conexión directa FC

No es posible conectar directamente un host a un sistema de almacenamiento ONTAP mediante el protocolo FC. La razón es el uso de NPIV. El WWN que identifica un puerto ONTAP FC con la red de FC utiliza un tipo de virtualización denominado NPIV. Cualquier dispositivo conectado a un sistema ONTAP debe poder reconocer un WWN de NPIV. No hay proveedores de HBA actuales que ofrezcan un HBA que se pueda instalar en un host que admita un destino NPIV.

Configuración de red en sistemas ASA r2

Interfaces lógicas

Las bases de datos de Oracle necesitan acceder al almacenamiento. Las interfaces lógicas (LIF) son las tuberías de red que conecta una máquina virtual de almacenamiento (SVM) a la red y, por tanto, a la base de datos. Es necesario diseñar un LIF adecuado para garantizar que exista un ancho de banda suficiente para cada carga de trabajo de la base de datos. La conmutación por error no conlleva la pérdida de los servicios de almacenamiento.

Esta sección proporciona una descripción general de los principios de diseño de LIF clave para sistemas ASA r2, que están optimizados para entornos exclusivamente SAN. Para obtener documentación más completa, consulte la "Documentación de gestión de red de ONTAP". Al igual que con otros aspectos de la arquitectura de bases de datos, las mejores opciones para el diseño de máquinas virtuales de almacenamiento (SVM, conocidas como vserver en la CLI) e interfaces lógicas (LIF) dependen en gran medida de los requisitos de escalabilidad y las necesidades comerciales.

Tenga en cuenta los siguientes temas principales al crear una estrategia de LIF:

- **Actuación.** ¿El ancho de banda de la red es suficiente para las cargas de trabajo de Oracle?
- **Resiliencia.** ¿Hay algún punto de falla en el diseño?
- **Capacidad de gestión.** ¿Se puede escalar la red de forma no disruptiva?

Estos temas se aplican a la solución completa, desde el host, los switches y el sistema de almacenamiento.

Tipos de LIF

Hay varios tipos de LIF. "Documentación de ONTAP sobre tipos de LIF" Puede proporcionar información más completa sobre este tema, pero desde una perspectiva funcional, los LIF se pueden dividir en los siguientes grupos:

- **LIF de administración de clúster y nodos.** LIF utilizadas para administrar el clúster de almacenamiento.
- **LIF de administración de SVM.** Interfaces que permiten el acceso a una SVM a través de la API REST o ONTAPI (también conocida como ZAPI) para funciones como la creación de instantáneas o el redimensionamiento de volúmenes. Productos como SnapManager para Oracle (SMO) deben tener acceso a una LIF de gestión de SVM.
- **LIF de datos.** Interfaces solo para protocolos SAN: FC, iSCSI, NVMe/FC, NVMe/TCP. Los protocolos NAS (NFS, SMB/CIFS) no son compatibles con los sistemas ASA r2.

> No es posible configurar una interfaz tanto para iSCSI (o NVMe/TCP) como para el tráfico de administración, a pesar de que ambos utilizan un protocolo IP. Se requiere un LIF de administración independiente en entornos iSCSI o NVMe/TCP. Para lograr resiliencia y rendimiento, configure múltiples LIF de datos SAN por protocolo por nodo y distribúyalos entre diferentes puertos físicos y estructuras. A diferencia de los sistemas AFF/ FAS , ASA r2 no permite tráfico NFS o SMB, por lo que no hay opción de reutilizar un LIF de datos NAS para su administración.

Diseño de LIF SAN

El diseño de LIF en un entorno SAN es relativamente sencillo por una de las razones: La multivía. Todas las implementaciones de SAN modernas permiten a un cliente acceder a los datos a través de múltiples rutas de red independientes y seleccionar la mejor ruta o las mejores rutas para acceder. Como resultado, el rendimiento con respecto al diseño de las LIF es más sencillo de abordar porque los clientes SAN equilibran automáticamente la carga de I/O en las mejores rutas disponibles.

Si una ruta deja de estar disponible, el cliente selecciona automáticamente una ruta diferente. La simplicidad resultante del diseño hace que los LIF SAN sean generalmente más gestionables. Esto no significa que un entorno SAN siempre se gestione con mayor facilidad, ya que existen otros muchos aspectos del almacenamiento SAN que son mucho más complicados que NFS. Simplemente significa que el diseño de LIF SAN es más sencillo.

Rendimiento

La consideración más importante con el rendimiento de LIF en un entorno SAN es el ancho de banda. Por ejemplo, un clúster ASA r2 de dos nodos con dos puertos FC de 32 Gb por nodo permite hasta 64 Gb de ancho de banda hacia/desde cada nodo. De manera similar, para NVMe/TCP o iSCSI, asegúrese de tener suficiente conectividad de 25 GbE o 100 GbE para las cargas de trabajo de Oracle.

Resiliencia

Los LIF SAN no se conmutan por error de la misma manera que los LIF NAS. Los sistemas ASA r2 dependen de la multirruta de host (MPIO/ALUA) para garantizar su resiliencia. Si un SAN LIF deja de estar disponible debido a una conmutación por error del controlador, el software de múltiples rutas del cliente detecta la pérdida de una ruta y redirige la E/S a una ruta alternativa. ASA r2 puede realizar una reubicación de LIF después de una breve demora para restaurar la disponibilidad total de la ruta, pero esto no interrumpe la E/S porque ya existen rutas activas en el nodo asociado. El proceso de conmutación por error se produce para restaurar el acceso del host en todos los puertos definidos.

Gran capacidad de administración

No es necesario migrar un LIF en un entorno SAN cuando los volúmenes se reubican dentro del par HA. Esto se debe a que, una vez completado el movimiento del volumen, ONTAP envía una notificación a la SAN sobre un cambio en las rutas y los clientes SAN vuelven a optimizar automáticamente. La migración de LIF con SAN se asocia principalmente con cambios importantes de hardware físico. Por ejemplo, si se requiere una actualización no disruptiva de los controladores, se migra una SAN LIF al nuevo hardware. Si se detecta que un puerto FC presenta fallas, se puede migrar un LIF a un puerto no utilizado.

Recomendaciones de diseño

NetApp hace las siguientes recomendaciones para entornos SAN ASA r2:

- No cree más rutas de las necesarias. Un número excesivo de rutas complica la gestión general y puede provocar problemas en la conmutación al nodo de respaldo de rutas en algunos hosts. Además, algunos hosts tienen limitaciones inesperadas de la ruta para configuraciones como el arranque SAN.
- Muy pocas configuraciones deberían requerir más de cuatro rutas a una LUN. El valor de tener más de dos nodos de rutas de publicidad para los LUN es limitado porque no se puede acceder al agregado que aloja una LUN si se produce un error en el nodo propietario de la LUN y su partner de alta disponibilidad. La creación de rutas en nodos que no sean el par de alta disponibilidad primario no es útil en esta situación.
- Aunque puede gestionar el número de rutas visibles de LUN si selecciona qué puertos se incluyen en las zonas FC, suele ser más fácil incluir todos los puntos de destino potenciales en la zona FC y controlar la

visibilidad de la LUN a nivel de ONTAP.

- Utilice la función de mapeo selectivo de LUN (SLM), que está habilitada de forma predeterminada. Con SLM, cualquier LUN nuevo se anuncia automáticamente desde el nodo que posee el agregado subyacente y el socio HA del nodo. Esta disposición evita la necesidad de crear conjuntos de puertos o configurar zonificación para limitar la accesibilidad del puerto. Cada LUN está disponible en la cantidad mínima de nodos necesarios para lograr un rendimiento y una resiliencia óptimos.
- En caso de que se deba migrar un LUN fuera de los dos controladores, se pueden agregar los nodos adicionales con el `lun mapping add-reporting-nodes` comando para que los LUN se anuncien en los nuevos nodos. Al hacerlo, se crean rutas SAN adicionales a los LUN para la migración de LUN. Sin embargo, el host debe realizar una operación de descubrimiento para utilizar las nuevas rutas.
- No se preocupe demasiado por el tráfico indirecto. Es mejor evitar el tráfico indirecto en un entorno con un gran volumen de I/O para el que cada microsegundo de latencia es crucial, pero el efecto de rendimiento visible es insignificante para las cargas de trabajo típicas.

Configuración TCP/IP y ethernet

Muchos clientes de Oracle on ASA r2 ONTAP utilizan Ethernet, el protocolo de red de iSCSI y NVMe/TCP.

Configuración del sistema operativo host

La mayoría de la documentación del proveedor de aplicaciones incluye configuraciones TCP y ethernet específicas para garantizar que la aplicación funcione de manera óptima. Estas mismas configuraciones suelen ser suficientes para ofrecer también un rendimiento óptimo del almacenamiento basado en IP.

Control de flujo Ethernet

Esta tecnología permite a un cliente solicitar que un remitente detenga temporalmente la transmisión de datos. Esto suele hacerse porque el receptor no puede procesar los datos entrantes con la suficiente rapidez. Al mismo tiempo, solicitar que un remitente cesara la transmisión era menos perjudicial que tener un receptor descarte de paquetes porque los buffers estaban llenos. Este ya no es el caso con las pilas TCP utilizadas en los sistemas operativos actualmente. De hecho, el control de flujo causa más problemas de los que resuelve.

Los problemas de rendimiento causados por el control de flujo de Ethernet han aumentado en los últimos años. Esto se debe a que el control de flujo Ethernet funciona en la capa Physical. Si una configuración de red permite que un sistema operativo del host envíe una solicitud de control de flujo de Ethernet a un sistema de almacenamiento, el resultado es una pausa en las operaciones de I/O de todos los clientes conectados. Debido a que una única controladora de almacenamiento atiende cada vez más a un número de clientes, la probabilidad de que uno o varios de estos clientes envíen solicitudes de control de flujo aumenta. El problema se ha observado con frecuencia en las instalaciones de los clientes con una amplia virtualización del SO.

Una NIC de un sistema NetApp no debe recibir solicitudes de control de flujo. El método utilizado para lograr este resultado varía según el fabricante del conmutador de red. En la mayoría de los casos, el control de flujo en un conmutador Ethernet se puede establecer en `receive desired` o. `receive on`, lo que significa que una solicitud de control de flujo no se reenvía al controlador de almacenamiento. En otros casos, la conexión de red en la controladora de almacenamiento puede no permitir la deshabilitación de control de flujo. En estos casos, los clientes deben configurarse para que nunca envíen solicitudes de control de flujo, ya sea cambiando a la configuración de NIC en el propio servidor host o a los puertos de switch a los que está conectado el servidor host.

Para los sistemas ASA r2, que son solo SAN, las consideraciones de control de flujo de Ethernet se aplican principalmente al tráfico iSCSI y NVMe/TCP.

* NetApp recomienda* asegurarse de que los controladores de almacenamiento NetApp ASA r2 no reciban paquetes de control de flujo Ethernet. Generalmente, esto se puede hacer configurando los puertos del conmutador a los que está conectado el controlador, pero algunos componentes del conmutador tienen limitaciones que podrían requerir cambios del lado del cliente.

Tamaños de MTU

Se ha demostrado que el uso de tramas gigantes ofrece alguna mejora del rendimiento en las redes 1GB al reducir la sobrecarga de la CPU y de la red, pero el beneficio no suele ser significativo.

NetApp recomienda implementar marcos jumbo cuando sea posible, tanto para obtener beneficios potenciales de rendimiento como para preparar la solución para el futuro.

Para los sistemas ASA r2, que son solo SAN, las tramas gigantes se aplican solo a protocolos SAN basados en Ethernet (iSCSI y NVMe/TCP).

El uso de tramas gigantes en una red 10Gb es casi obligatorio. Esto se debe a que la mayoría de las implementaciones de 10Gb alcanzan un límite de paquetes por segundo sin tramas gigantes antes de alcanzar la marca de 10Gb. El uso de tramas gigantes mejora la eficiencia del procesamiento TCP/IP porque permite que el sistema operativo, el servidor, las NIC y el sistema de almacenamiento procesen menos paquetes, pero más grandes. La mejora del rendimiento varía de NIC a NIC, pero es significativa.

En las implementaciones de tramas gigantes, existe la creencia común, aunque incorrecta, de que todos los dispositivos conectados deben admitir tramas gigantes y que el tamaño de MTU debe coincidir de extremo a extremo En su lugar, los dos extremos de red negocian el tamaño de trama más alto mutuamente aceptable al establecer una conexión. En un entorno típico, un switch de red se establece con un tamaño de MTU de 9216, la controladora NetApp se establece en 9000 y los clientes se configuran con una combinación de 9000 y 1514. Los clientes que admiten un MTU de 9000 pueden utilizar tramas gigantes, y los clientes que solo puedan admitir 1514 pueden negociar un valor inferior.

Los problemas con esta disposición son raros en un entorno completamente conmutado. Sin embargo, tenga cuidado en un entorno enrutado que ningún enrutador intermedio se vea forzado a fragmentar tramas gigantes.

- NetApp recomienda* configurar lo siguiente para entornos SAN ASA r2:
- Los marcos gigantes son deseables pero no obligatorios con 1 GbE.
- Se requieren tramas gigantes para obtener el máximo rendimiento con 10 GbE y más rápido para el tráfico iSCSI y NVMe/TCP.

Parámetros de TCP

A menudo hay tres ajustes mal configurados: Marcas de tiempo TCP, reconocimiento selectivo (SACK) y escalado de ventana TCP. Muchos documentos desactualizados en Internet recomiendan deshabilitar uno o varios de estos parámetros para mejorar el rendimiento. Había algo de mérito en esta recomendación hace muchos años, cuando las capacidades de la CPU eran mucho menores y había un beneficio en reducir la sobrecarga en el procesamiento TCP siempre que fuera posible.

Sin embargo, con los sistemas operativos modernos, deshabilitar cualquiera de estas características de TCP generalmente no resulta en ningún beneficio detectable, a la vez que también puede dañar el rendimiento. Los daños en el rendimiento son especialmente probables en entornos de red virtualizados, ya que estas características son necesarias para gestionar eficazmente la pérdida de paquetes y los cambios en la calidad

de la red.

NetApp recomienda habilitar las marcas de tiempo TCP, EL SACK y el escalado de la ventana TCP en el host, y los tres parámetros deben estar activados por defecto en cualquier sistema operativo actual.

Configuración de SAN FC

La configuración de FC SAN para bases de datos Oracle en sistemas ASA r2 se trata principalmente de seguir las mejores prácticas de SAN estándar.

ASA r2 está optimizado para cargas de trabajo exclusivas de SAN, por lo que los principios siguen siendo los mismos que los de AFF/ FAS, con un enfoque en el rendimiento, la resiliencia y la simplicidad. Esto incluye medidas de planificación típicas, como garantizar que exista suficiente ancho de banda en la SAN entre el host y el sistema de almacenamiento, verificar que existan todas las rutas SAN entre todos los dispositivos requeridos, usar las configuraciones de puerto FC requeridas por su proveedor de conmutador FC, evitar la contención de ISL y usar un monitoreo adecuado de la estructura SAN.

División en zonas

Una zona de FC nunca debe contener más de un iniciador. Tal arreglo puede parecer funcionar inicialmente, pero la comunicación entre iniciadores finalmente interfiere con el rendimiento y la estabilidad.

Las zonas multidestino se consideran generalmente seguras, aunque en raras ocasiones el comportamiento de los puertos de destino FC de diferentes proveedores ha causado problemas. Por ejemplo, evite incluir los puertos de destino de una cabina de almacenamiento NetApp y otra que no sea de NetApp en la misma zona. Además, es aún más probable que la ubicación de un sistema de almacenamiento NetApp y un dispositivo de cinta en la misma zona cause problemas.

- ASA r2 utiliza zonas de disponibilidad de almacenamiento en lugar de agregados, pero esto no cambia los principios de zonificación de FC.
- La multirruta (MPIO) sigue siendo el principal mecanismo de resiliencia; sin embargo, para los sistemas ASA r2 que admiten multirruta activa-activa simétrica, todas las rutas a un LUN están activas y se utilizan para E/S simultáneamente.

Red de conexión directa

A veces, los administradores de almacenamiento prefieren simplificar sus infraestructuras eliminando los switches de red de la configuración. Esto puede ser soportado en algunos escenarios.

ISCSI y NVMe/TCP

Un host que utiliza iSCSI o NVMe/TCP puede conectarse directamente a un sistema de almacenamiento ASA r2 y funcionar normalmente. La razón es el pathing. Las conexiones directas a dos controladores de almacenamiento diferentes dan como resultado dos rutas independientes para el flujo de datos. La pérdida de ruta, puerto o controlador no impide que se utilice la otra ruta, siempre que la función de rutas múltiples esté configurada correctamente.

Conexión directa FC

No es posible conectar directamente un host a un sistema de almacenamiento ASA r2 utilizando el protocolo FC. La razón es la misma que con los sistemas AFF/ FAS , uso de NPIV. El WWN que identifica un puerto FC de ONTAP a la red FC utiliza un tipo de virtualización llamado NPIV. Cualquier dispositivo conectado a un sistema ONTAP debe poder reconocer un WWN NPIV. No hay proveedores de HBA actuales que ofrezcan un HBA que pueda instalarse en un host que pueda soportar un objetivo NPIV.

Configuración del almacenamiento en sistemas AFF/FAS

FC SAN

Alineación de LUN

La alineación de LUN hace referencia a optimizar las I/O con respecto al diseño del sistema de archivos subyacente.

En un sistema ONTAP, el almacenamiento se organiza en 4KB unidades. Un bloque 8KB de base de datos o sistema de archivos debe asignarse exactamente a dos bloques de 4KB KB. Si un error de configuración de una LUN cambia la alineación 1KB en cualquier dirección, cada bloque de 8KB KB existiría en tres bloques de almacenamiento de 4KB KB diferentes en lugar de dos. Esta disposición provocaría una mayor latencia y provocaría la realización de I/O adicionales en el sistema de almacenamiento.

La alineación también afecta a las arquitecturas LVM. Si se define un volumen físico de un grupo de volúmenes lógicos en todo el dispositivo de la unidad (no se crean particiones), el primer bloque de 4KB KB del LUN se alinea con el primer bloque de 4KB KB del sistema de almacenamiento. Esta es una alineación correcta. Los problemas surgen con las particiones porque cambian la ubicación inicial en la que el sistema operativo utiliza la LUN. Siempre que la compensación se desplaza en unidades enteras de 4KB, la LUN se alinea.

En entornos Linux, cree grupos de volúmenes lógicos en todo el dispositivo de la unidad. Cuando se necesita una partición, compruebe la alineación ejecutando `fdisk -u` y verificando que el inicio de cada partición es un múltiplo de ocho. Esto significa que la partición comienza en un múltiplo de ocho sectores de 512 bytes, que es 4KB.

Consulte también la discusión sobre la alineación de los bloques de compresión en la sección "Eficiencia". Cualquier diseño alineado con los límites de bloques de compresión de 8KB KB también se alineará con los límites de 4KB KB.

Advertencias de desalineación

El registro de rehacer/transacciones de bases de datos normalmente genera I/O no alineadas que pueden provocar advertencias engañosas acerca de las LUN mal alineadas en ONTAP.

El registro realiza una escritura secuencial del archivo log con escrituras de tamaño variable. Una operación de escritura de registro que no se alinea con los límites de 4KB no provoca problemas de rendimiento normalmente, ya que la próxima operación de escritura de registro completa el bloque. El resultado es que ONTAP es capaz de procesar casi todas las escrituras de bloques de 4KB KB completos, aunque los datos de algunos bloques de 4KB KB se hayan escrito en dos operaciones independientes.

Verifique la alineación mediante el uso de utilidades como `sio` o. `dd` Que puede generar I/O en un tamaño de bloque definido. Las estadísticas de alineación de I/O del sistema de almacenamiento se pueden ver con `stats` comando. Consulte "Verificación de la alineación de WAFL" si quiere más información.

La alineación en entornos Solaris es más complicada. Consulte "Configuración de host SAN ONTAP" si quiere más información.

Precaución
En entornos Solaris x86, tenga cuidado adicional con la alineación correcta, ya que la mayoría de las configuraciones tienen varias capas de particiones. Los segmentos de partición de Solaris x86 normalmente existen en la parte superior de una tabla de particiones de registro de inicio maestro estándar.

Ajuste de tamaño de LUN y número de LUN

Seleccionar el tamaño óptimo de LUN y el número de LUN que se utilizarán es fundamental para lograr un rendimiento y una capacidad de gestión óptimos en las bases de datos de Oracle.

Una LUN es un objeto virtualizado en ONTAP que existe en todas las unidades del agregado host. Como resultado, el rendimiento de la LUN no se ve afectado por su tamaño porque la LUN aprovecha todo el potencial de rendimiento del agregado sin importar el tamaño que se haya elegido.

Para comodidad, es posible que los clientes deseen usar una LUN de un tamaño determinado. Por ejemplo, si una base de datos se crea en un LVM u un grupo de discos de ASM de Oracle compuesto por dos LUN de 1TB GB cada uno, dicho grupo de discos debe aumentar en incrementos de 1TB TB. Es preferible crear el grupo de discos a partir de ocho LUN de 500GB cada uno para que el grupo de discos se pueda aumentar en incrementos menores.

Se desaconseja la práctica de establecer un tamaño de LUN estándar universal porque, al hacerlo, se puede complicar la capacidad de gestión. Por ejemplo, un tamaño de LUN estándar de 100GB TB puede funcionar bien cuando una base de datos o un almacén de datos está entre 1TB y 2TB TB, pero el tamaño de una base de datos o un almacén de datos de 20TB TB requeriría 200 LUN. Esto significa que los tiempos de reinicio del servidor son más largos, hay más objetos que gestionar en las distintas interfaces de usuario y productos como SnapCenter deben realizar la detección de muchos objetos. Si se usa menos LUN, de mayor tamaño se evitan estos problemas.

- El número de LUN es más importante que el tamaño de la LUN.
- El tamaño de LUN está controlado principalmente por requisitos del número de LUN.
- Evite crear más LUN de las necesarias.

Número de LUN

A diferencia del tamaño de LUN, el número de LUN afecta al rendimiento. El rendimiento de la aplicación depende a menudo de la capacidad para realizar I/O paralelas mediante la capa SCSI. Como resultado, dos LUN ofrecen mejor rendimiento que una única LUN. El uso de LVM como Veritas VxVM, Linux LVM2 u Oracle ASM es el método más sencillo para aumentar el paralelismo.

Los clientes de NetApp suelen experimentar un beneficio mínimo gracias al aumento del número de LUN por encima de dieciséis, aunque, en pruebas de entornos 100% con unidades de estado sólido con I/O aleatorias muy pesadas, se ha demostrado una mejora adicional de hasta 64 000 LUN.

NetApp recomienda lo siguiente:

En general, entre cuatro y dieciséis LUN son suficientes para admitir las necesidades de I/O de cualquier carga de trabajo de bases de datos en concreto. Menos de cuatro LUN puede crear limitaciones de rendimiento debido a las limitaciones de las implementaciones SCSI del host.

Ubicación de LUN

La colocación óptima de los LUN de bases de datos en volúmenes de ONTAP depende principalmente de cómo se utilicen varias funciones de ONTAP.

Volúmenes

Un punto común de confusión con los clientes que empiezan a utilizar ONTAP es el uso de FlexVols, conocido normalmente como «volúmenes».

Un volumen no es una LUN. Estos términos se usan sinónimos con muchos otros productos de proveedores, incluidos los proveedores de cloud. Los volúmenes de ONTAP son simplemente contenedores de gestión. No sirven datos por sí mismas, ni ocupan el espacio. Son contenedores para archivos o LUN y existen para mejorar y simplificar la capacidad de gestión, especialmente a escala.

Volúmenes y LUN

Normalmente, los LUN relacionados se ubican en un único volumen. Por ejemplo, una base de datos que requiere 10 LUN suele tener 10 LUN colocadas en el mismo volumen.

- Usar una proporción 1:1 de LUN y volúmenes, lo que significa una LUN por volumen, no es * una práctica recomendada formal.
- En su lugar, los volúmenes deben verse como contenedores para las cargas de trabajo o conjuntos de datos. Puede que haya una única LUN por volumen, o que haya muchos. La respuesta correcta depende de los requisitos de capacidad de gestión.
- La dispersión de LUN por un número innecesario de volúmenes puede provocar problemas de sobrecarga adicionales y programación para operaciones como las operaciones de snapshot, el número excesivo de objetos que se muestran en la interfaz de usuario y que pueda alcanzar los límites de volúmenes de plataforma antes de alcanzar el límite de LUN.

Volúmenes, LUN y snapshots

Las políticas y las programaciones de Snapshot se colocan en el volumen, no en la LUN. Un conjunto de datos formado por 10 LUN solo requeriría una única política de Snapshot cuando esas LUN se ubiquen en el mismo volumen.

Además, la coubicación de todas las LUN relacionadas para un conjunto de datos determinado en un único volumen proporciona operaciones de instantánea atómica. Por ejemplo, una base de datos que residía en 10 LUN o un entorno de aplicación basado en VMware formado por 10 sistemas operativos diferentes podría protegerse como un único objeto consistente si las LUN subyacentes se colocan en un único volumen. Si se colocan en diferentes volúmenes, las instantáneas pueden o no estar sincronizadas al 100%, incluso si se programan al mismo tiempo.

En algunos casos, podría haber que dividir un conjunto relacionado de LUN en dos volúmenes distintos debido a los requisitos de recuperación. Por ejemplo, una base de datos podría tener cuatro LUN para archivos de datos y dos LUN para registros. En este caso, un volumen de archivo de datos con 4 LUN y un volumen de registro con 2 LUN podrían ser la mejor opción. La razón es la capacidad de recuperación independiente. Por ejemplo, el volumen de archivos de datos se podría restaurar de forma selectiva a un estado anterior, lo que significa que las cuatro LUN se revertirían al estado de la snapshot, mientras que el volumen de registro con sus datos cruciales no se vería afectado.

Volúmenes, LUN y SnapMirror

Las políticas y las operaciones de SnapMirror son, como las operaciones de Snapshot, realizadas en el volumen, no en la LUN.

Ubicar conjuntamente LUN relacionadas en un único volumen le permite crear una única relación de SnapMirror y actualizar todos los datos contenidos con una única actualización. Al igual que con las instantáneas, la actualización también será una operación atómica. Se garantizaría que el destino de SnapMirror tendrá una única réplica puntual de los LUN de origen. Si las LUN se distribuyeron entre varios volúmenes, las réplicas pueden o no ser coherentes entre sí.

Volúmenes, LUN y calidad de servicio

Aunque la calidad de servicio se puede aplicar de forma selectiva a LUN individuales, normalmente es más fácil configurarla en el nivel de volumen. Por ejemplo, todas las LUN utilizadas por los invitados de un servidor ESX determinado podrían colocarse en un solo volumen y, a continuación, podría aplicarse una política de calidad de servicio adaptable de ONTAP. El resultado es un límite IOPS por TB con escala automática que se aplica a todas las LUN.

Del mismo modo, si una base de datos necesitara 100K 000 IOPS y ocupase 10 LUN, sería más fácil establecer un único límite de 100K IOPS en un único volumen que establecer 10 límites individuales de 10K IOPS, uno en cada LUN.

Diseños de varios volúmenes

Hay algunos casos en los que distribuir las LUN en varios volúmenes puede ser beneficioso. El motivo primario es la segmentación de la controladora. Por ejemplo, un sistema de almacenamiento de alta disponibilidad podría estar alojando una única base de datos donde se requiera todo el potencial de procesamiento y almacenamiento en caché de cada controladora. En este caso, un diseño típico sería colocar la mitad de las LUN de un único volumen de la controladora 1 y la otra mitad de los LUN de un único volumen en la controladora 2.

Del mismo modo, la segmentación de la controladora puede utilizarse para equilibrar la carga. Un sistema de alta disponibilidad que alojara 100 bases de datos de 10 LUN cada una se podría diseñar donde cada base de datos reciba un volumen de 5 LUN en cada una de las dos controladoras. El resultado es una carga simétrica garantizada de cada controladora a medida que se aprovisionan las bases de datos adicionales.

Sin embargo, ninguno de estos ejemplos implica una relación de volumen/LUN de 1:1 GB. El objetivo sigue siendo optimizar la gestión mediante la colocalización de LUN relacionadas en volúmenes.

Un ejemplo donde tiene sentido la relación de 1:1 LUN con volumen es la colocación en contenedores, donde cada LUN podría representar realmente una única carga de trabajo y cada una de ellas debería gestionarse de forma individual. En tales casos, una relación 1:1 puede ser óptima.

Cambio de tamaño de LUN y cambio de tamaño de LVM

Cuando un sistema de archivos basado en SAN ha alcanzado su límite de capacidad, hay dos opciones para aumentar el espacio disponible:

- Aumente el tamaño de las LUN
- Agregue una LUN a un grupo de volúmenes existente y aumente el volumen lógico contenido

Aunque el redimensionamiento de LUN es una opción para aumentar la capacidad, generalmente es mejor usar un LVM, incluido Oracle ASM. Uno de los principales motivos por los que existen LVM es evitar la necesidad de cambiar el tamaño de las LUN. Con un LVM, se unen varias LUN en un pool virtual de

almacenamiento. Los volúmenes lógicos tallados en este pool son administrados por el LVM y pueden ser fácilmente redimensionados. Otra ventaja es la eliminación de los puntos de sobrecarga en una unidad concreta al distribuir un volumen lógico determinado entre todas las LUN disponibles. Normalmente, la migración transparente puede realizarse utilizando el administrador de volúmenes para reubicar las extensiones subyacentes de un volumen lógico a nuevas LUN.

Segmentación de LVM

La segmentación de LVM hace referencia a distribuir datos entre varias LUN. El resultado es una mejora espectacular del rendimiento en muchas bases de datos.

Antes de la era de las unidades flash, se utilizaba la segmentación para ayudar a superar las limitaciones de rendimiento de las unidades giratorias. Por ejemplo, si un sistema operativo necesita realizar una operación de lectura de 1MB KB, para leer que 1MB TB de datos de una sola unidad se requeriría buscar y leer muchos cabezales de unidad ya que 1MB se transfiere lentamente. Si esos 1MB TB de datos se segmentaron en 8 LUN, el sistema operativo podría emitir ocho operaciones de lectura de 128K KB en paralelo y reducir el tiempo necesario para realizar la transferencia de 1MB GB.

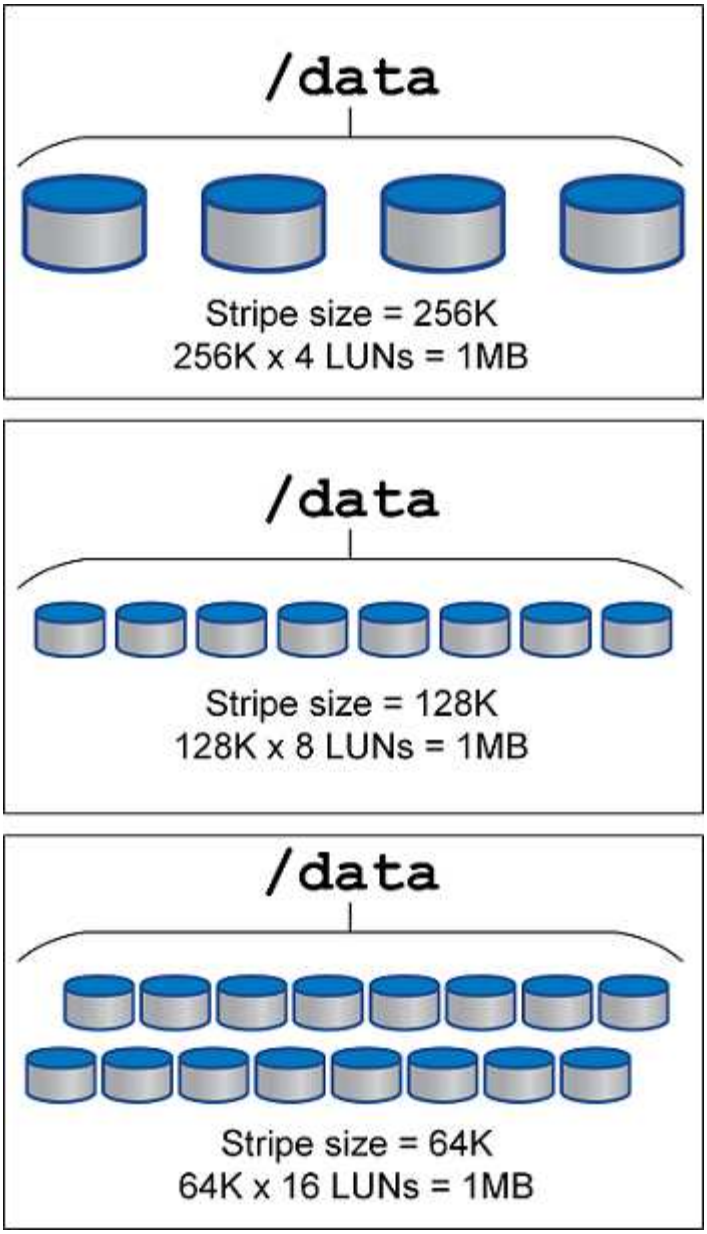
La segmentación con unidades giratorias era más difícil porque se tenía que conocer el patrón de I/O con anterioridad. Si la segmentación no se ajustó correctamente para los patrones de I/O reales, las configuraciones seccionadas podrían dañar el rendimiento. Con las bases de datos de Oracle y, especialmente con las configuraciones all-flash, la segmentación es mucho más fácil de configurar y se ha demostrado que mejora drásticamente el rendimiento.

Los gestores de volúmenes lógicos como Oracle ASM segmentan por defecto, pero el LVM del sistema operativo nativo no lo hacen. Algunos de ellos unen varias LUN como un dispositivo concatenado, lo que da como resultado archivos de datos que existen en un único dispositivo LUN. Esto provoca puntos calientes. Otras implementaciones de LVM toman por defecto extensiones distribuidas. Esto es similar a la segmentación, pero es más grueso. Las LUN del grupo de volúmenes se dividen en partes grandes, denominadas extensiones y normalmente se miden en muchos megabytes, y los volúmenes lógicos se distribuyen por esas extensiones. El resultado es que las operaciones de I/O aleatorias en un archivo se deben distribuir bien entre las LUN, pero las operaciones de I/O secuenciales no son tan eficientes como podrían.

La I/O de aplicaciones con rendimiento intensivo casi siempre es una (a) en unidades del tamaño de bloque básico o (b) un megabyte.

El principal objetivo de una configuración seccionada es garantizar que la I/O de archivo único se pueda realizar como una unidad única y que las I/O de varios bloques, que deben tener un tamaño de 1MB TB, se puedan paralelizar de manera uniforme entre todas las LUN del volumen seccionado. Esto significa que el tamaño de franja no debe ser menor que el tamaño del bloque de la base de datos y el tamaño de franja multiplicado por el número de LUN debe ser 1MB.

En la siguiente figura, se muestran tres opciones posibles para el ajuste del tamaño de la franja y el ancho. Se selecciona el número de LUN para satisfacer los requisitos de rendimiento tal como se han descrito anteriormente, pero en todos los casos los datos totales de una sola franja es 1MB.



NFS

Descripción general

NetApp lleva más de 30 años proporcionando almacenamiento NFS de clase empresarial y su uso está creciendo con la tendencia hacia las infraestructuras basadas en cloud debido a la sencillez de la tecnología.

El protocolo NFS incluye varias versiones con diferentes requisitos. Para obtener una descripción completa de la configuración de NFS con ONTAP, consulte "TR-4067 NFS en prácticas recomendadas de ONTAP". Las siguientes secciones cubren algunos de los requisitos más críticos y los errores comunes del usuario.

Versiones de NFS

El cliente NFS del sistema operativo debe ser compatible con NetApp.

- NFSv3 es compatible con sistemas operativos que siguen el estándar NFSv3.

- NFSv3 es compatible con el cliente Oracle dNFS.
- NFSv4 es compatible con todos los sistemas operativos que siguen el estándar NFSv4.
- NFSv4,1 y NFSv4,2 requieren soporte de SO específico. Consulte la "NetApp IMT" Para sistemas operativos compatibles.
- La compatibilidad de Oracle dNFS para NFSv4,1 requiere Oracle 12.2.0.2 o superior.

> La "Matriz de compatibilidad de NetApp" Para NFSv3 y NFSv4 no incluye sistemas operativos específicos. Todos los sistemas operativos que obedecen a RFC son generalmente compatibles. Al buscar en IMT en línea compatibilidad con NFSv3 o NFSv4, no seleccione un sistema operativo concreto porque no se mostrarán coincidencias. Todos los sistemas operativos están soportados implícitamente por la política general.

Tablas de ranuras TCP Linux NFSv3

Las tablas de ranuras TCP son equivalentes a NFSv3 a la profundidad de la cola del adaptador de bus de host (HBA). En estas tablas se controla el número de operaciones de NFS que pueden extraordinarias a la vez. El valor predeterminado suele ser 16, que es demasiado bajo para un rendimiento óptimo. El problema opuesto ocurre en los kernels más nuevos de Linux, que pueden aumentar automáticamente el límite de la tabla de ranuras TCP a un nivel que sature el servidor NFS con solicitudes.

Para obtener un rendimiento óptimo y evitar problemas de rendimiento, ajuste los parámetros del núcleo que controlan las tablas de ranuras TCP.

Ejecute el `sysctl -a | grep tcp.*.slot_table` command, y observe los siguientes parámetros:

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

Todos los sistemas Linux deben incluir `sunrpc.tcp_slot_table_entries`, pero solo algunos incluyen `sunrpc.tcp_max_slot_table_entries`. Ambos deben establecerse en 128.

> Si no se establecen estos parámetros, puede tener efectos significativos en el rendimiento. En algunos casos, el rendimiento es limitado porque el sistema operativo linux no está emitiendo suficiente I/O. En otros casos, las latencias de I/O aumentan cuando el sistema operativo linux intenta emitir más operaciones de I/O de las que se pueden mantener.

ADR y NFS

Algunos clientes han informado de problemas de rendimiento derivados de una cantidad excesiva de I/O en los datos de `ADR` ubicación. Por lo general, el problema no ocurre hasta que se acumulan muchos datos de rendimiento. Se desconoce el motivo del exceso de E/S, pero este problema parece ser el resultado de que los procesos de Oracle exploran repetidamente el directorio de destino en busca de cambios.

Extracción del `noac` y/o. `actimeo=0` Las opciones de montaje permiten almacenar en caché el sistema operativo del host y reducen los niveles de I/O de almacenamiento.

NetApp recomienda no colocar `ADR` datos en un sistema de archivos con `noac` o. `actimeo=0` ya que son probables problemas de rendimiento. Separar `ADR` los datos en un punto de montaje diferente si es necesario.

nfs-rootonly y mount-rootonly

ONTAP incluye una opción de NFS denominada `nfs-rootonly` Esto controla si el servidor acepta conexiones de tráfico NFS desde puertos altos. Como medida de seguridad, solo el usuario root puede abrir conexiones TCP/IP utilizando un puerto de origen inferior a 1024 porque dichos puertos normalmente están reservados para el uso del sistema operativo, no para los procesos del usuario. Esta restricción ayuda a garantizar que el tráfico NFS provenga de un cliente NFS del sistema operativo real y no de un proceso malicioso que emula un cliente NFS. El cliente dNFS de Oracle es un controlador de espacio de usuario, pero el proceso se ejecuta como raíz, por lo que generalmente no es necesario cambiar el valor de `nfs-rootonly`. Las conexiones se realizan a partir de puertos bajos.

La `mount-rootonly` La opción solo se aplica a NFSv3. Controla si la llamada DE MONTAJE RPC se acepta desde puertos superiores a 1024. Cuando se utiliza dNFS, el cliente vuelve a ejecutarse como raíz, por lo que puede abrir puertos por debajo de 1024. Este parámetro no tiene efecto.

Los procesos que abren conexiones con dNFS a través de NFS versiones 4,0 y superiores no se ejecutan como raíz y, por lo tanto, requieren puertos a través de 1024. La `nfs-rootonly` El parámetro debe estar establecido en disabled para que dNFS complete la conexión.

Si `nfs-rootonly` Está habilitada, el resultado es un bloqueo durante la fase de montaje al abrir las conexiones dNFS. La salida sqlplus tiene un aspecto similar al siguiente:

```
SQL>startup
ORACLE instance started.
Total System Global Area 4294963272 bytes
Fixed Size                  8904776 bytes
Variable Size             822083584 bytes
Database Buffers         3456106496 bytes
Redo Buffers                7868416 bytes
```

El parámetro se puede cambiar de la siguiente manera:

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

En raras ocasiones, es posible que necesite cambiar nfs-rootonly y mount-rootonly a disabled. Si un servidor administra un número extremadamente grande de conexiones TCP, es posible que no haya puertos por debajo de 1024 GbE disponibles y que el sistema operativo se vea forzado a utilizar puertos más altos. Estos dos parámetros de ONTAP necesitarían ser cambiados para permitir que la conexión se complete.

Políticas de exportación NFS: Superusuario y setuid

Si los binarios de Oracle se encuentran en un recurso compartido NFS, la política de exportación debe incluir permisos de superusuario y setuid.

Las exportaciones NFS compartidas que se utilizan para servicios de archivos genéricos, como los directorios iniciales de usuario, suelen aplastar al usuario raíz. Esto significa que una solicitud del usuario root en un host que ha montado un sistema de archivos se vuelve a asignar como un usuario diferente con privilegios inferiores. Esto ayuda a proteger los datos al impedir que un usuario root de un servidor determinado acceda a los datos del servidor compartido. El bit setuid también puede ser un riesgo de seguridad en un entorno compartido. El bit setuid permite que un proceso se ejecute como un usuario diferente al usuario que llama al comando. Por ejemplo, un script de shell que era propiedad de root con el bit setuid se ejecuta como root. Si ese script de shell pudiera ser cambiado por otros usuarios, cualquier usuario que no sea root podría emitir un comando como root actualizando el script.

Los binarios de Oracle incluyen archivos propiedad de root y utilizan el bit setuid. Si los binarios de Oracle se instalan en un recurso compartido NFS, la política de exportación debe incluir los permisos de superusuario y setuid adecuados. En el ejemplo siguiente, la regla incluye ambos `allow-suid` y permisos `superuser` Acceso (root) para clientes NFS mediante la autenticación del sistema.

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

Configuración de NFSv4/4,1

Para la mayoría de las aplicaciones, hay muy poca diferencia entre NFSv3 y NFSv4. Las operaciones de I/O de aplicaciones suelen ser muy sencillas y no se benefician de forma significativa de algunas de las funciones avanzadas disponibles en NFSv4. Las versiones superiores de NFS no deberían considerarse como una «actualización» desde el punto de vista del almacenamiento de base de datos, sino como versiones de NFS que incluyen funciones adicionales. Por ejemplo, si se requiere la seguridad de extremo a extremo del modo de privacidad de kerberos (krb5p), se necesita NFSv4.

NetApp recomienda usar NFSv4,1 si se requieren capacidades de NFSv4. Existen algunas mejoras funcionales en el protocolo NFSv4 en NFSv4,1 que mejoran la resiliencia en ciertos casos perimetrales.

Cambiar a NFSv4 es más complicado que simplemente cambiar las opciones de montaje de vers=3 a vers=4,1. Para obtener una explicación más completa de la configuración de NFSv4 con ONTAP, que incluye instrucciones para configurar el sistema operativo, consulte "Prácticas recomendadas de TR-4067 NFS en ONTAP". En las siguientes secciones de este documento técnico se explican algunos de los requisitos básicos para el uso de NFSv4.

NFSv4 dominio

Una explicación completa de la configuración de NFSv4/4,1 está fuera del alcance de este documento, pero un problema que se encuentra comúnmente es una discrepancia en la asignación de dominio. Desde un punto de vista sysadmin, los sistemas de archivos NFS parecen comportarse normalmente, pero las aplicaciones informan de errores sobre permisos y/o setuid en determinados archivos. En algunos casos, los administradores han concluido incorrectamente que los permisos de los binarios de la aplicación se han dañado y han ejecutado comandos chown o chmod cuando el problema real era el nombre de dominio.

El nombre de dominio NFSv4 se establece en la SVM de ONTAP:

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

El nombre de dominio NFSv4 del host se establece en `/etc/idmap.cfg`

```
[root@host1 etc]# head /etc/idmapd.conf
[General]
#Verbosity = 0
# The following should be set to the local NFSv4 domain name
# The default is the host's DNS domain name.
Domain = my.lab
```

Los nombres de dominio deben coincidir. Si no lo hacen, aparecerán errores de asignación similares a los siguientes en la `/var/log/messages`:

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

Los binarios de aplicaciones, como los binarios de Oracle Database, incluyen archivos propiedad de root con el bit setuid, lo que significa que una discrepancia en los nombres de dominio NFSv4 provoca fallos en el inicio de Oracle y una advertencia sobre la propiedad o los permisos de un archivo llamado `oradism`, que se encuentra en la `$ORACLE_HOME/bin` directorio. Debería aparecer de la siguiente manera:

```
[root@host1 etc]# ls -l /orabin/product/19.3.0.0/dbhome_1/bin/oradism
-rwsr-x--- 1 root oinstall 147848 Apr 17  2019
/orabin/product/19.3.0.0/dbhome_1/bin/oradism
```

Si este archivo aparece con la propiedad de Nadie, puede haber un problema de asignación de dominio NFSv4.

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

Para solucionarlo, compruebe la `/etc/idmap.cfg` Haga un archivo con la configuración de v4-id-domain en ONTAP y asegúrese de que son consistentes. Si no lo son, realice los cambios necesarios, ejecute `nfsidmap -c`, y esperar un momento para que los cambios se propaguen. La propiedad del archivo debe reconocerse correctamente como root. Si un usuario había intentado ejecutar `chown root` En este archivo antes de que se corrigiera la configuración de los dominios NFS, es posible que sea necesario ejecutarlo `chown root` de nuevo.

NFS directo de Oracle (dNFS)

Las bases de datos de Oracle pueden utilizar NFS de dos maneras.

En primer lugar, puede utilizar un sistema de archivos montado utilizando el cliente NFS nativo que forma parte del sistema operativo. A veces, esto se denomina nfs del núcleo o knfs. El sistema de archivos NFS es montado y utilizado por la base de datos Oracle exactamente igual que cualquier otra aplicación utilizaría un sistema de archivos NFS.

El segundo método es Oracle Direct NFS (dNFS). Se trata de una implementación del estándar NFS dentro del software de base de datos Oracle. No cambia la forma en que el DBA configura o gestiona las bases de datos Oracle. Siempre que el sistema de almacenamiento disponga de la configuración correcta, el uso de dNFS debe ser transparente para el equipo de administradores de bases de datos y los usuarios finales.

Una base de datos con la función dNFS activada todavía tiene montados los sistemas de archivos NFS habituales. Una vez abierta la base de datos, Oracle Database abre un conjunto de sesiones TCP/IP y ejecuta las operaciones NFS directamente.

NFS directo

El valor principal de Direct NFS de Oracle es omitir el cliente NFS host y realizar operaciones de archivos NFS directamente en un servidor NFS. Para activarlo sólo es necesario cambiar la biblioteca de Oracle Disk Manager (ODM). Las instrucciones para este proceso se proporcionan en la documentación de Oracle.

El uso de dNFS permite mejorar considerablemente el rendimiento de I/O y disminuye la carga en el host y el sistema de almacenamiento, ya que el proceso de I/O se realiza de la forma más eficiente posible.

Además, Oracle dNFS incluye una **opción** para el acceso múltiple de la interfaz de red y la tolerancia a fallos. Por ejemplo, se pueden enlazar dos interfaces de 10Gb GbE para ofrecer un ancho de banda de 20Gb Gb/s. El fallo de una interfaz provoca que se vuelvan a intentar I/O en la otra interfaz. El funcionamiento general es muy similar al multivía FC. La tecnología MultiPath era común hace años, cuando ethernet de 1GB Gb era el estándar más común. Una NIC de 10Gb es suficiente para la mayoría de las cargas de trabajo de Oracle, pero si se necesitan más, se pueden vincular 10Gb NIC.

Cuando se utiliza dNFS, es crítico que se instalen todos los parches descritos en Oracle Doc 1495104,1. Si no se puede instalar un parche, se debe evaluar el entorno para asegurarse de que los errores descritos en ese documento no causen problemas. En algunos casos, la imposibilidad de instalar los parches necesarios impide el uso de dNFS.

No utilice dNFS con ningún tipo de resolución de nombres por turnos, incluidos DNS, DDNS, NIS o cualquier otro método. Esto incluye la función de equilibrio de carga DNS disponible en ONTAP. Cuando una base de datos Oracle que utiliza dNFS resuelve un nombre de host en una dirección IP, no debe cambiar en las consultas posteriores. Esto puede provocar fallos en la base de datos de Oracle y daños en los datos.

Habilitando dNFS

Oracle dNFS puede trabajar con NFSv3 sin necesidad de configuración más allá de habilitar la biblioteca dNFS (consulte la documentación de Oracle para ver el comando específico necesario), pero si dNFS no puede establecer la conectividad, puede volver silenciosamente al cliente NFS del núcleo. Si esto sucede, el rendimiento puede verse seriamente afectado.

Si desea utilizar la multiplexación dNFS a través de múltiples interfaces, con NFSv4.X, o utilizar el cifrado, debe configurar un archivo oranfstab. La sintaxis es extremadamente estricta. Pequeños errores en el archivo pueden provocar el bloqueo del inicio o el paso por alto del archivo oranfstab.

En el momento de la redacción de este documento, la función multivía de dNFS no funciona con NFSv4,1 con las versiones recientes de Oracle Database. Un archivo oranfstab que especifica NFSv4,1 como protocolo sólo puede utilizar una sentencia PATH única para una exportación determinada. El motivo es que ONTAP no admite el trunking ClientID. Los parches de Oracle Database para resolver esta limitación pueden estar disponibles en el futuro.

La única forma de estar seguro de que dNFS funciona como se espera es consultar las tablas v$dnfs.

A continuación se muestra un archivo oranfstab de ejemplo ubicado en /etc. Esta es una de las múltiples ubicaciones en las que se puede colocar un archivo oranfstab.

```
[root@jfs11 trace]# cat /etc/oranfstab
server: NFSv3test
path: jfs_svmdr-nfs1
path: jfs_svmdr-nfs2
export: /dbf mount: /oradata
export: /logs mount: /logs
nfs_version: NFSv3
```

El primer paso es comprobar que dNFS está operativo para los sistemas de archivos especificados:

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

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

/dbf
NFSv3.0
```

Esta salida indica que dNFS está en uso con estos dos sistemas de archivos, pero **no** significa que oranfstab está operativo. Si se presentara un error, dNFS habría detectado automáticamente los sistemas de archivos NFS del host y es posible que todavía vea la misma salida de este comando.

La multivía se puede comprobar de la siguiente manera:

```
SQL> select svrname,path,ch_id from v$dnfs_channels;

SVRNAME
-----------------------------------
PATH
-----------------------------------
     CH_ID
```

```
----------
NFSv3test
jfs_svmdr-nfs1
         0

NFSv3test
jfs_svmdr-nfs2
         1

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------

NFSv3test
jfs_svmdr-nfs1
         0

NFSv3test
jfs_svmdr-nfs2

[output truncated]

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------
NFSv3test
jfs_svmdr-nfs2
         1

NFSv3test
jfs_svmdr-nfs1
         0

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------
```

```
NFSv3test
jfs_svmdr-nfs2
         1


66 rows selected.
```

Estas son las conexiones que utiliza dNFS. Hay dos rutas y canales visibles para cada entrada SVRNAME. Esto significa que la multivía está funcionando, lo que significa que se reconoció y procesó el archivo oranfstab.

Acceso directo a sistemas de archivos del host y NFS

En ocasiones, el uso de dNFS puede ocasionar problemas en las aplicaciones o actividades del usuario que se basan en los sistemas de archivos visibles montados en el host, ya que el cliente dNFS accede al sistema de archivos fuera de banda desde el sistema operativo host. El cliente dNFS puede crear, eliminar y modificar archivos sin el conocimiento del sistema operativo.

Cuando se utilizan las opciones de montaje para bases de datos de instancia única, se activa el almacenamiento en caché de atributos de archivo y directorio, lo que también significa que el contenido de un directorio está en caché. Por lo tanto, dNFS puede crear un archivo, y hay un breve retraso antes de que el sistema operativo vuelva a leer el contenido del directorio y el archivo se haga visible para el usuario. Esto no es generalmente un problema, pero, en raras ocasiones, utilidades como SAP BR*Tools pueden tener problemas. Si esto sucede, solucione el problema cambiando las opciones de montaje para utilizar las recomendaciones para Oracle RAC. Este cambio provoca la deshabilitación de todo el almacenamiento en caché del host.

Cambie las opciones de montaje solo cuando (a) se utiliza dNFS y (b) se produce un problema debido a un desfase en la visibilidad de los archivos. Si no se utiliza dNFS, el rendimiento se reduce al utilizar las opciones de montaje de Oracle RAC en una base de datos de instancia única.

> Consulte la nota `nosharecache` sobre en "Opciones de montaje de Linux NFS" para ver un problema de dNFS específico de Linux que puede producir resultados inusuales.

Arrendamientos y bloqueos de NFS

NFSv3 está sin estado. Esto implica efectivamente que el servidor NFS (ONTAP) no realiza un seguimiento de qué sistemas de archivos están montados, quién o qué bloqueos están realmente instalados.

ONTAP dispone de algunas funciones que registrarán los intentos de montaje, por lo que tiene una idea de qué clientes pueden acceder a los datos y puede que haya bloqueos asesores, pero no se garantiza que esa información esté al 100% completa. No se puede completar, ya que el seguimiento del estado del cliente NFS no forma parte del estándar NFSv3.1.

NFSv4 Estado

Por el contrario, NFSv4 tiene estado. El servidor NFSv4 rastrea qué clientes están utilizando qué sistemas de archivos, qué archivos existen, qué archivos y/o regiones de archivos están bloqueados, etc. Esto significa que debe haber una comunicación regular entre un servidor NFSv4 para mantener los datos de estado actualizados.

Los estados más importantes que gestiona el servidor NFS son NFSv4 bloqueos y NFSv4 arrendamientos y están muy entrelazados. Necesitas entender cómo cada uno trabaja por sí mismo, y cómo se relacionan entre sí.

NFSv4 bloqueos

Con NFSv3, las cerraduras son un aviso. Un cliente NFS aún puede modificar o eliminar un archivo «bloqueado». Un bloqueo NFSv3 no caduca por sí mismo, debe ser eliminado. Esto crea problemas. Por ejemplo, si tiene una aplicación en cluster que crea NFSv3 bloqueos y uno de los nodos falla, ¿qué debe hacer? Puede codificar la aplicación en los nodos supervivientes para eliminar los bloqueos, pero ¿cómo sabe que es seguro? ¿Puede que el nodo «fallido» esté operativo, pero no se comunica con el resto del clúster?

Con NFSv4, las cerraduras tienen una duración limitada. Mientras el cliente que mantiene los bloqueos continúe registrando en el servidor NFSv4, no se permitirá a ningún otro cliente adquirir estos bloqueos. Si un cliente no se registra en NFSv4, el servidor eventualmente revoca los bloqueos y otros clientes podrán solicitar y obtener bloqueos.

NFSv4 arrendamientos

NFSv4 bloqueos están asociados a un arrendamiento NFSv4. Cuando un cliente NFSv4 establece una conexión con un servidor NFSv4, obtiene un permiso. Si el cliente obtiene un bloqueo (hay muchos tipos de bloqueos), el bloqueo se asocia con la concesión.

Esta concesión tiene un timeout definido. De forma predeterminada, ONTAP establecerá el valor de tiempo de espera en 30 segundos:

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

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

Esto significa que un cliente NFSv4 necesita registrarse con el servidor NFSv4 cada 30 segundos para renovar sus arrendamientos.

El arrendamiento se renueva automáticamente por cualquier actividad, por lo que si el cliente está haciendo trabajo no hay necesidad de realizar operaciones de adición. Si una aplicación se vuelve silenciosa y no está haciendo un trabajo real, tendrá que realizar una especie de operación de mantenimiento de la vida (llamada SECUENCIA) en su lugar. En esencia, es solo decir «sigo aquí, actualice mis contratos de arrendamiento».

```
 *Question:* What happens if you lose network connectivity for 31 seconds?
NFSv3 está sin estado. No se espera la comunicación de los clientes. NFSv4
aparece con estado y, una vez que transcurre el período de concesión, la
concesión caduca, se revocan los bloqueos, y los archivos bloqueados se
ponen a disposición de otros clientes.
```

Con NFSv3, puede mover los cables de red, reiniciar los switches de red, realizar cambios de configuración y estar bastante seguro de que no sucedería nada malo. Las aplicaciones normalmente solo esperarían pacientemente a que la conexión de red vuelva a funcionar.

Con NFSv4, tienes 30 segundos (a menos que hayas aumentado el valor de ese parámetro dentro de ONTAP) para completar tu trabajo. Si sobrepasa eso, se agota el tiempo de arrendamiento. Normalmente, esto provoca fallos de aplicación.

Por ejemplo, si tiene una base de datos Oracle y experimenta una pérdida de conectividad de red (a veces denominada «partición de red») que supera el tiempo de espera de concesión, bloqueará la base de datos.

A continuación, se muestra un ejemplo de lo que ocurre en el log de alertas de Oracle si esto sucede:

```
2022-10-11T15:52:55.206231-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00202: control file: '/redo0/NTAP/ctrl/control01.ctl'
ORA-27072: File I/O error
Linux-x86_64 Error: 5: Input/output error
Additional information: 4
Additional information: 1
Additional information: 4294967295
2022-10-11T15:52:59.842508-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00206: error in writing (block 3, # blocks 1) of control file
ORA-00202: control file: '/redo1/NTAP/ctrl/control02.ctl'
ORA-27061: waiting for async I/Os failed
```

Si observa los syslogs, debería ver varios de estos errores:

```
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
```

Los mensajes de registro suelen ser el primer signo de un problema, aparte de la congelación de la aplicación. Normalmente, no verá nada durante la interrupción de la red, porque los procesos y el propio SO están bloqueados al intentar acceder al sistema de archivos NFS.

Los errores aparecen después de que la red vuelva a funcionar. En el ejemplo anterior, una vez que se restablece la conectividad, el sistema operativo intentó volver a adquirir los bloqueos, pero era demasiado tarde. El arrendamiento había caducado y se eliminaron los bloqueos. Esto produce un error que se propaga hasta la capa de Oracle y provoca el mensaje en el log de alertas. Es posible que vea variaciones en estos patrones en función de la versión y la configuración de la base de datos.

En resumen, NFSv3 tolera la interrupción de la red, pero NFSv4 es más sensible e impone un período de arrendamiento definido.

¿Qué pasa si un tiempo de espera de 30 segundos no es aceptable? ¿Qué pasa si administra una red que cambia dinámicamente en la que se reinician los switches o se reubican los cables y el resultado es la interrupción ocasional de la red? Puede optar por ampliar el período de arrendamiento, pero si lo desea,

requiere una explicación de NFSv4 períodos de gracia.

NFSv4 periodos de gracia

Si se reinicia un servidor NFSv3, está listo para servir IO casi al instante. No estaba manteniendo ningún tipo de estado sobre los clientes. El resultado es que la operación de toma de control de ONTAP parece estar casi al instante. En el momento en que un controlador está listo para comenzar a servir datos, enviará un ARP a la red que indica el cambio en la topología. En general, los clientes lo detectan de forma casi instantánea y se reanuda el flujo de los datos.

NFSv4, sin embargo, producirá una breve pausa. Es solo parte de cómo funciona NFSv4.

> Las siguientes secciones están actualizadas a partir de ONTAP 9.15,1, pero el comportamiento de arrendamiento y bloqueo, así como las opciones de ajuste pueden cambiar de versión a versión. Si necesita ajustar los tiempos de espera de arrendamiento/bloqueo NFSv4, consulte al soporte de NetApp para obtener la información más reciente.

Los servidores NFSv4 necesitan realizar un seguimiento de los arrendamientos, los bloqueos y quién utiliza qué datos. Si un servidor NFS produce una alarma y se reinicia, pierde energía durante un momento, o se reinicia durante la actividad de mantenimiento, el resultado es la concesión/bloqueo y se pierde otra información del cliente. El servidor necesita averiguar qué cliente está utilizando qué datos antes de reanudar las operaciones. Aquí es donde entra el período de gracia.

Si de repente apaga el servidor NFSv4. Cuando vuelva a estar activo, los clientes que intenten reanudar I/O obtendrán una respuesta que diga «He perdido información de arrendamiento/bloqueo. ¿Desea volver a registrar sus bloqueos? Ese es el comienzo del período de gracia. El valor predeterminado es 45 segundos en ONTAP:

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

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

El resultado es que, después de un reinicio, una controladora pausará el I/O mientras todos los clientes recuperan sus concesiones y bloqueos. Una vez que finaliza el período de gracia, el servidor reanudará las operaciones de E/S.

Este período de gracia controla la reclamación de concesión durante los cambios de la interfaz de red, pero hay un segundo período de gracia que controla la recuperación durante la conmutación por error del almacenamiento, `locking.grace_lease_seconds`. Esta es una opción de nivel de nodo.

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

Por ejemplo, si necesitaba realizar recuperaciones tras fallos de LIF y necesitaba reducir el período de gracia, cambiaría `v4-grace-seconds`. Si desea mejorar el tiempo de reanudación de I/O durante la recuperación tras fallos de la controladora, debería cambiar `locking.grace_lease_seconds`.

Solo altere estos valores con precaución y después de comprender completamente los riesgos y las

consecuencias. Las pausas de I/O implicadas en las operaciones de recuperación tras fallos y migración con NFSv4.X no se pueden evitar completamente. Los períodos de bloqueo, arrendamiento y gracia forman parte de NFS RFC. Para muchos clientes, NFSv3 es preferible porque los tiempos de recuperación tras fallos son más rápidos.

Tiempos de espera de leasing frente a períodos de gracia

El período de gracia y el período de arrendamiento están conectados. Como se ha mencionado anteriormente, el tiempo de espera predeterminado de la concesión es de 30 segundos, lo que significa que NFSv4 clientes deben realizar el check in con el servidor al menos cada 30 segundos o pierden sus arrendamientos y, a su vez, sus bloqueos. El período de gracia existe para permitir que un servidor NFS vuelva a generar los datos de concesión/bloqueo y, de forma predeterminada, es de 45 segundos. El período de gracia debe ser superior al período de arrendamiento. Esto garantiza que un entorno de cliente NFS diseñado para renovar arrendamientos al menos cada 30 segundos pueda conectarse con el servidor después de un reinicio. Un período de gracia de 45 segundos asegura que todos aquellos clientes que esperan renovar sus arrendamientos al menos cada 30 segundos definitivamente tienen la oportunidad de hacerlo.

Si un tiempo de espera de 30 segundos no es aceptable, puede optar por ampliar el período de arrendamiento.

Si desea aumentar el tiempo de espera de concesión a 60 segundos para soportar una interrupción de la red de 60 segundos, también tendrá que aumentar el período de gracia. Esto significa que experimentará pausas más largas de I/O durante la recuperación tras fallos de la controladora.

Esto no debería ser normalmente un problema. Los usuarios habituales solo actualizan las controladoras de ONTAP una o dos veces al año, y las recuperaciones tras fallos no planificadas debido a fallos de hardware son extremadamente raras. Además, si tenía una red en la que una interrupción de la red de 60 segundos era preocupante y necesitaba un tiempo de espera de concesión de 60 segundos, es probable que no se oponga a una conmutación por error rara del sistema de almacenamiento, lo que provoca una pausa de 61 segundos. Ya ha reconocido que tiene una red que se detiene durante más de 60 segundos con bastante frecuencia.

Almacenamiento en caché NFS

La presencia de cualquiera de las siguientes opciones de montaje provoca la deshabilitación del almacenamiento en caché del host:

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

Estos ajustes pueden tener un efecto negativo grave en la velocidad de la instalación del software, la aplicación de parches y las operaciones de copia de seguridad/restauración. En algunos casos, especialmente con aplicaciones en cluster, estas opciones son necesarias como consecuencia inevitable de la necesidad de proporcionar coherencia de la caché en todos los nodos del cluster. En otros casos, los clientes utilizan estos parámetros por error y el resultado es un daño innecesario en el rendimiento.

Muchos clientes eliminan temporalmente estas opciones de montaje durante la instalación o aplicación de parches de los archivos binarios de la aplicación. Esta eliminación se puede realizar de forma segura si el usuario comprueba que ningún otro proceso está utilizando activamente el directorio de destino durante el proceso de instalación o aplicación de parches.

Los tamaños de transferencia de NFS

De forma predeterminada, ONTAP limita el tamaño de I/O de NFS a 64K.

La I/O aleatoria con la mayoría de aplicaciones y bases de datos utiliza un tamaño de bloque mucho más pequeño, que es muy inferior al máximo de 64K KB. Las operaciones de I/O de grandes bloques suelen estar en paralelo, por lo que el máximo de 64K KB tampoco se limita a obtener el ancho de banda máximo.

Hay algunas cargas de trabajo en las que el máximo de 64K crea una limitación. En particular, las operaciones de subproceso único como la operación de copia de seguridad o recuperación o una exploración de tabla completa de la base de datos se ejecutan de forma más rápida y eficiente si la base de datos puede realizar menos E/S pero más grandes. El tamaño óptimo de gestión de I/O para ONTAP es de 256K KB.

El tamaño de transferencia máximo para una SVM de ONTAP determinada se puede cambiar de la siguiente manera:

```
Cluster01::> set advanced
Warning: These advanced commands are potentially dangerous; use them only
when directed to do so by NetApp personnel.
Do you want to continue? {y|n}: y
Cluster01::*> nfs server modify -vserver vserver1 -tcp-max-xfer-size
262144
Cluster01::*>
```

> No reduzca nunca el tamaño máximo permitido de transferencia en ONTAP por debajo del valor de rsize/wsize de los sistemas de archivos NFS montados actualmente. Esto puede crear bloqueos o incluso corrupción de datos con algunos sistemas operativos. Por ejemplo, si los clientes NFS se establecen actualmente con un valor de rsize/wsize de 65536 000, el tamaño de transferencia máximo de ONTAP se podría ajustar entre 65536 000 y 1048576 000 sin que ello afecte a porque los propios clientes están limitados. Reducir el tamaño máximo de transferencia por debajo de 65536 puede dañar la disponibilidad o los datos.

NVFAIL

NVFAIL es una función de ONTAP que garantiza la integridad en situaciones catastróficas de conmutación por error.

Las bases de datos son vulnerables a daños durante eventos de conmutación por error de almacenamiento debido a que mantienen cachés internos de gran tamaño. Si un evento catastrófico requiere forzar una conmutación por error de ONTAP o forzar la conmutación por error de MetroCluster, independientemente del estado de la configuración general, el resultado es que los cambios confirmados previamente se pueden descartar de forma efectiva. El contenido de la cabina de almacenamiento se retrocede en el tiempo y el estado de la caché de base de datos ya no refleja el estado de los datos del disco. Esta inconsistencia provoca daños en los datos.

El almacenamiento en caché puede tener lugar en la capa de aplicaciones o del servidor. Por ejemplo, una configuración de Oracle Real Application Cluster (RAC) con servidores activos tanto en un sitio primario como en un sitio remoto almacena datos en caché en Oracle SGA. Una operación de conmutación de sitios forzada que provocara una pérdida de datos pondría la base de datos en riesgo de dañarse, ya que los bloques almacenados en el SGA podrían no coincidir con los bloques del disco.

Un uso menos obvio del almacenamiento en caché se da en la capa del sistema de archivos del sistema de sistemas operativos. Los bloques de un sistema de archivos NFS montado se pueden almacenar en caché en el sistema operativo. Como alternativa, un sistema de archivos en clúster basado en las LUN ubicadas en el sitio primario podría montarse en servidores en el sitio remoto y una vez más podrían almacenarse los datos

en caché. Un fallo de NVRAM o una toma de control forzada o una conmutación de sitios forzada en estas situaciones podría provocar daños en el sistema de archivos.

ONTAP protege las bases de datos y los sistemas operativos de este escenario con NVFAIL y su configuración asociada.

Utilidad de Reclamación de ASM (ASMRU)

ONTAP elimina de manera eficiente los bloques puestos a cero que se escriben en un archivo o LUN cuando se habilita la compresión en línea. Las utilidades como Oracle ASM Reclamation Utility (ASRU) funcionan escribiendo ceros en extensiones de ASM no utilizadas.

Esto permite a los administradores de bases de datos reclamar espacio en la cabina de almacenamiento después de la eliminación de los datos. ONTAP intercepta los ceros y desasigna el espacio de la LUN. El proceso de recuperación es extremadamente rápido porque no se escriben datos en el sistema de almacenamiento.

Desde el punto de vista de la base de datos, el grupo de discos de ASM contiene ceros, y leer esas regiones de las LUN daría como resultado un flujo de ceros, pero ONTAP no almacena los ceros en las unidades. En su lugar, se realizan cambios sencillos en los metadatos que marcan internamente las regiones en cero de la LUN como vacías de datos.

Por motivos similares, las pruebas de rendimiento que involucran datos puestos a cero no son válidas porque en realidad los bloques de ceros no se procesan como escrituras en la cabina de almacenamiento.

> Al utilizar ASRU, asegúrese de que todos los parches recomendados por Oracle están instalados.

Configuración de almacenamiento en sistemas ASA R2

FC SAN

Alineación de LUN

La alineación de LUN hace referencia a optimizar las I/O con respecto al diseño del sistema de archivos subyacente.

Los sistemas ASA r2 utilizan la misma arquitectura ONTAP que AFF/ FAS pero con un modelo de configuración simplificado. Los sistemas ASA r2 utilizan zonas de disponibilidad de almacenamiento (SAZ) en lugar de agregados, pero los principios de alineación siguen siendo los mismos porque ONTAP administra el diseño de bloques de manera consistente en todas las plataformas. Sin embargo, tenga en cuenta estos puntos específicos de ASA:

- Los sistemas ASA r2 proporcionan rutas simétricas activas-activas para todos los LUN, lo que elimina los problemas de asimetría de rutas durante la alineación.
- Las unidades de almacenamiento (LUN) tienen aprovisionamiento fino de manera predeterminada; la alineación no cambia este comportamiento.
- La reserva de instantáneas y la eliminación automática de instantáneas se pueden configurar durante la creación de LUN (ONTAP 9.18.1 y posteriores).

En un sistema ONTAP, el almacenamiento se organiza en 4KB unidades. Un bloque 8KB de base de datos o sistema de archivos debe asignarse exactamente a dos bloques de 4KB KB. Si un error de configuración de una LUN cambia la alineación 1KB en cualquier dirección, cada bloque de 8KB KB existiría en tres bloques de almacenamiento de 4KB KB diferentes en lugar de dos. Esta disposición provocaría una mayor latencia y provocaría la realización de I/O adicionales en el sistema de almacenamiento.

La alineación también afecta a las arquitecturas LVM. Si se define un volumen físico de un grupo de volúmenes lógicos en todo el dispositivo de la unidad (no se crean particiones), el primer bloque de 4KB KB del LUN se alinea con el primer bloque de 4KB KB del sistema de almacenamiento. Esta es una alineación correcta. Los problemas surgen con las particiones porque cambian la ubicación inicial en la que el sistema operativo utiliza la LUN. Siempre que la compensación se desplaza en unidades enteras de 4KB, la LUN se alinea.

En entornos Linux, cree grupos de volúmenes lógicos en todo el dispositivo de la unidad. Cuando se necesita una partición, compruebe la alineación ejecutando `fdisk -u` y verificando que el inicio de cada partición es un múltiplo de ocho. Esto significa que la partición comienza en un múltiplo de ocho sectores de 512 bytes, que es 4KB.

Consulte también la discusión sobre la alineación de los bloques de compresión en la sección "Eficiencia". Cualquier diseño alineado con los límites de bloques de compresión de 8KB KB también se alineará con los límites de 4KB KB.

Advertencias de desalineación

El registro de rehacer/transacciones de bases de datos normalmente genera I/O no alineadas que pueden provocar advertencias engañosas acerca de las LUN mal alineadas en ONTAP.

El registro realiza una escritura secuencial del archivo log con escrituras de tamaño variable. Una operación de escritura de registro que no se alinea con los límites de 4KB no provoca problemas de rendimiento normalmente, ya que la próxima operación de escritura de registro completa el bloque. El resultado es que ONTAP es capaz de procesar casi todas las escrituras de bloques de 4KB KB completos, aunque los datos de algunos bloques de 4KB KB se hayan escrito en dos operaciones independientes.

Verifique la alineación utilizando utilidades como `sio` o `dd` que puede generar E/S en un tamaño de bloque definido. Las estadísticas de alineación de E/S en el sistema de almacenamiento se pueden ver con el `stats` dominio. Ver "Verificación de la alineación de WAFL" Para más información.

La alineación en entornos Solaris es más complicada. Consulte "Configuración de host SAN ONTAP" si quiere más información.

> En entornos Solaris x86, tenga cuidado adicional con la alineación correcta, ya que la mayoría de las configuraciones tienen varias capas de particiones. Los segmentos de partición de Solaris x86 normalmente existen en la parte superior de una tabla de particiones de registro de inicio maestro estándar.

Mejores prácticas adicionales:

- Verifique el firmware de HBA y la configuración del sistema operativo con la herramienta Matriz de interoperabilidad de NetApp (IMT).
- Utilice las utilidades sanlun para confirmar el estado y la alineación de la ruta.
- Para Oracle ASM y LVM, asegúrese de que los archivos de configuración (/etc/lvm/lvm.conf, /etc/sysconfig/oracleasm) estén configurados correctamente para evitar problemas de alineación.

Ajuste de tamaño de LUN y número de LUN

Seleccionar el tamaño óptimo de LUN y el número de LUN que se utilizarán es fundamental para lograr un rendimiento y una capacidad de gestión óptimos en las bases de datos de Oracle.

Un LUN es un objeto virtualizado en ONTAP que existe en todas las unidades de la zona de disponibilidad de almacenamiento (SAZ) de alojamiento en los sistemas ASA r2. Como resultado, el rendimiento del LUN no se ve afectado por su tamaño porque el LUN aprovecha todo el potencial de rendimiento del SAZ sin importar el tamaño elegido.

Para comodidad, es posible que los clientes deseen usar una LUN de un tamaño determinado. Por ejemplo, si una base de datos se crea en un LVM u un grupo de discos de ASM de Oracle compuesto por dos LUN de 1TB GB cada uno, dicho grupo de discos debe aumentar en incrementos de 1TB TB. Es preferible crear el grupo de discos a partir de ocho LUN de 500GB cada uno para que el grupo de discos se pueda aumentar en incrementos menores.

Se desaconseja la práctica de establecer un tamaño de LUN estándar universal porque, al hacerlo, se puede complicar la capacidad de gestión. Por ejemplo, un tamaño de LUN estándar de 100GB TB puede funcionar bien cuando una base de datos o un almacén de datos está entre 1TB y 2TB TB, pero el tamaño de una base de datos o un almacén de datos de 20TB TB requeriría 200 LUN. Esto significa que los tiempos de reinicio del servidor son más largos, hay más objetos que gestionar en las distintas interfaces de usuario y productos como SnapCenter deben realizar la detección de muchos objetos. Si se usa menos LUN, de mayor tamaño se evitan estos problemas.

- Consideraciones ASA r2: *
- El tamaño máximo de LUN para ASA r2 es 128 TB, lo que permite menos LUN más grandes sin afectar el rendimiento.
- ASA r2 utiliza zonas de disponibilidad de almacenamiento (SAZ) en lugar de agregados, pero esto no cambia la lógica de tamaño de LUN para las cargas de trabajo de Oracle.
- El aprovisionamiento fino está habilitado de manera predeterminada; cambiar el tamaño de los LUN no causa interrupciones y no requiere desconectarlos.

Número de LUN

A diferencia del tamaño de LUN, el número de LUN afecta al rendimiento. El rendimiento de la aplicación depende a menudo de la capacidad para realizar I/O paralelas mediante la capa SCSI. Como resultado, dos LUN ofrecen mejor rendimiento que una única LUN. El uso de LVM como Veritas VxVM, Linux LVM2 u Oracle ASM es el método más sencillo para aumentar el paralelismo.

Con ASA r2, los principios para el conteo de LUN siguen siendo los mismos que en AFF/ FAS porque ONTAP maneja la E/S paralela de manera similar en todas las plataformas. Sin embargo, la arquitectura solo SAN de ASA r2 y las rutas simétricas activas-activas garantizan un rendimiento consistente en todos los LUN.

Los clientes de NetApp suelen experimentar un beneficio mínimo gracias al aumento del número de LUN por encima de dieciséis, aunque, en pruebas de entornos 100% con unidades de estado sólido con I/O aleatorias muy pesadas, se ha demostrado una mejora adicional de hasta 64 000 LUN.

NetApp recomienda lo siguiente:

En general, de cuatro a dieciséis LUN son suficientes para satisfacer las necesidades de E/S de cualquier carga de trabajo de base de datos Oracle. Menos de cuatro LUN pueden crear limitaciones de rendimiento debido a limitaciones en las implementaciones SCSI del host. Aumentar más de dieciséis LUN rara vez mejora el rendimiento, excepto en casos extremos (como cargas de trabajo de SSD de E/S aleatorias muy altas).

Ubicación de LUN

La ubicación óptima de los LUN de base de datos dentro de los sistemas ASA r2 depende principalmente de cómo se utilizarán las distintas características de ONTAP .

En los sistemas ASA r2, las unidades de almacenamiento (LUN o espacios de nombres NVMe) se crean a partir de una capa de almacenamiento simplificada denominada Zonas de disponibilidad de almacenamiento (SAZ), que actúan como grupos comunes de almacenamiento para un par de alta disponibilidad.

Normalmente solo hay una zona de disponibilidad de almacenamiento (SAZ) por par de alta disponibilidad.

Zonas de disponibilidad de almacenamiento (SAZ)

En los sistemas ASA r2, los volúmenes siguen estando ahí, pero se crean automáticamente cuando se crean las unidades de almacenamiento. Las unidades de almacenamiento (LUN o espacios de nombres NVMe) se aprovisionan directamente dentro de los volúmenes creados automáticamente en las zonas de disponibilidad de almacenamiento (SAZ). Este diseño elimina la necesidad de gestión manual de volúmenes y hace que el aprovisionamiento sea más directo y optimizado para cargas de trabajo en bloque como las bases de datos Oracle.

Zonas de almacenamiento y unidades de almacenamiento

Las unidades de almacenamiento relacionadas (LUN o espacios de nombres NVMe) normalmente se ubican juntas dentro de una única zona de disponibilidad de almacenamiento (SAZ). Por ejemplo, una base de datos que requiere 10 unidades de almacenamiento (LUN) normalmente tendría las 10 unidades ubicadas en la misma SAZ para mayor simplicidad y rendimiento.

- El uso de una relación 1:1 de unidades de almacenamiento a volúmenes, es decir, una unidad de almacenamiento (LUN) por volumen, es el comportamiento predeterminado de ASA r2.
- En caso de que haya más de un par HA en el sistema ASA r2, las unidades de almacenamiento (LUN) para una base de datos determinada se pueden distribuir entre múltiples SAZ para optimizar el uso y el rendimiento del controlador.

En el contexto de FC SAN, aquí la unidad de almacenamiento se refiere a LUN.

Grupos de consistencia (CG), LUN e instantáneas

En ASA r2, las políticas y programaciones de instantáneas se aplican en el nivel de grupo de consistencia, que es una construcción lógica que agrupa múltiples LUN o espacios de nombres NVMe para una protección de datos coordinada. Un conjunto de datos que consta de 10 LUN solo requeriría una única política de instantánea cuando esos LUN sean parte del mismo grupo de consistencia.

Los grupos de consistencia garantizan operaciones de instantáneas atómicas en todos los LUN incluidos. Por ejemplo, una base de datos que reside en 10 LUN, o un entorno de aplicación basado en VMware que consta de 10 sistemas operativos diferentes, se puede proteger como un único objeto consistente si los LUN subyacentes se agrupan en el mismo grupo de consistencia. Si se colocan en diferentes grupos de consistencia, las instantáneas pueden o no estar perfectamente sincronizadas, incluso si se programan al mismo tiempo.

En algunos casos, puede ser necesario dividir un conjunto relacionado de LUN en dos grupos de consistencia diferentes debido a los requisitos de recuperación. Por ejemplo, una base de datos podría tener cuatro LUN para archivos de datos y dos LUN para registros. En este caso, un grupo de consistencia de archivos de datos con 4 LUN y un grupo de consistencia de registros con 2 LUN podrían ser la mejor opción. El motivo es la capacidad de recuperación independiente: el grupo de consistencia del archivo de datos podría restaurarse selectivamente a un estado anterior, lo que significa que los cuatro LUN volverían al estado de la instantánea, mientras que el grupo de consistencia del registro con sus datos críticos permanecería inafectado.

CG, LUN y SnapMirror

Las políticas y operaciones de SnapMirror se realizan, al igual que las operaciones de instantáneas, en el grupo de consistencia, no en el LUN.

La ubicación conjunta de LUN relacionados en un único grupo de consistencia le permite crear una única relación SnapMirror y actualizar todos los datos contenidos con una única actualización. Al igual que con las instantáneas, la actualización también será una operación atómica. Se garantizaría que el destino SnapMirror tenga una única réplica en un punto en el tiempo de los LUN de origen. Si los LUN se distribuyeron en múltiples grupos de consistencia, las réplicas podrían o no ser consistentes entre sí.

La replicación de SnapMirror en sistemas ASA r2 tiene las siguientes limitaciones:

- La replicación síncrona de SnapMirror no es compatible.
- La sincronización activa de SnapMirror solo es compatible entre dos sistemas ASA r2.
- La replicación asíncrona de SnapMirror solo es compatible entre dos sistemas ASA r2.
- La replicación asíncrona de SnapMirror no es compatible entre un sistema ASA r2 y un sistema ASA, AFF o FAS o la nube.

Obtenga más información sobre "Políticas de replicación de SnapMirror compatibles con sistemas ASA r2".

CG, LUN y QoS

Si bien la calidad de servicio se puede aplicar de forma selectiva a LUN individuales, generalmente es más fácil configurarla en el nivel del grupo de consistencia. Por ejemplo, todos los LUN utilizados por los invitados en un servidor ESX determinado podrían colocarse en un único grupo de consistencia y luego podría aplicarse una política de QoS adaptativa de ONTAP . El resultado es un límite de IOPS por TiB autoescalable que se aplica a todos los LUN.

De la misma manera, si una base de datos requiere 100 000 IOPS y ocupa 10 LUN, sería más fácil establecer un único límite de 100 000 IOPS en un único grupo de consistencia que establecer 10 límites individuales de 10 000 IOPS, uno en cada LUN.

Múltiples diseños CG

Hay algunos casos en los que puede resultar beneficioso distribuir LUN entre múltiples grupos de consistencia. La razón principal es la distribución del controlador. Por ejemplo, un sistema de almacenamiento

HA ASA r2 podría alojar una única base de datos Oracle donde se requiere todo el potencial de procesamiento y almacenamiento en caché de cada controlador. En este caso, un diseño típico sería colocar la mitad de los LUN en un solo grupo de consistencia en el controlador 1 y la otra mitad de los LUN en un solo grupo de consistencia en el controlador 2.

De manera similar, para entornos que alojan muchas bases de datos, la distribución de LUN en múltiples grupos de consistencia puede garantizar una utilización equilibrada del controlador. Por ejemplo, un sistema HA que aloja 100 bases de datos de 10 LUN cada una podría asignar 5 LUN a un grupo de consistencia en el controlador 1 y 5 LUN a un grupo de consistencia en el controlador 2 por base de datos. Esto garantiza una carga simétrica a medida que se aprovisionan bases de datos adicionales.

Sin embargo, ninguno de estos ejemplos implica una relación LUN-grupo de consistencia de 1:1. El objetivo sigue siendo optimizar la capacidad de administración agrupando los LUN relacionados de manera lógica en un grupo de consistencia.

Un ejemplo en el que una relación LUN a grupo de consistencia de 1:1 tiene sentido son las cargas de trabajo en contenedores, donde cada LUN puede representar en realidad una única carga de trabajo que requiere políticas de replicación e instantáneas independientes y, por lo tanto, deben gestionarse de forma individual. En tales casos, una proporción de 1:1 puede ser óptima.

Cambio de tamaño de LUN y cambio de tamaño de LVM

Cuando un sistema de archivos basado en SAN o un grupo de discos Oracle ASM alcanza su límite de capacidad en ASA r2, hay dos opciones para aumentar el espacio disponible:

- Aumentar el tamaño de las LUN existentes (unidades de almacenamiento)
- Agregue un nuevo LUN a un grupo de discos ASM o un grupo de volúmenes LVM existente y haga crecer el volumen lógico contenido

Aunque el cambio de tamaño de LUN es compatible con ASA r2, generalmente es mejor utilizar un administrador de volúmenes lógicos (LVM) como Oracle ASM. Una de las principales razones por las que existen los LVM es evitar la necesidad de cambiar el tamaño de los LUN con frecuencia. Con un LVM, varios LUN se combinan en un grupo virtual de almacenamiento. Los volúmenes lógicos extraídos de este grupo se pueden redimensionar fácilmente sin afectar la configuración de almacenamiento subyacente.

Los beneficios adicionales de usar LVM o ASM incluyen:

- Optimización del rendimiento: distribuye E/S entre múltiples LUN, lo que reduce los puntos críticos.
- Flexibilidad: agregue nuevos LUN sin interrumpir las cargas de trabajo existentes.
- Migración transparente: ASM o LVM pueden reubicar extensiones en nuevos LUN para equilibrar o organizar en niveles sin tiempo de inactividad del host.

Consideraciones clave de ASA r2:

- El cambio de tamaño de LUN se realiza a nivel de unidad de almacenamiento dentro de una máquina virtual de almacenamiento (SVM) utilizando la capacidad de la zona de disponibilidad de almacenamiento (SAZ).
- Para Oracle, la mejor práctica es agregar LUN a los grupos de discos ASM en lugar de cambiar el tamaño de los LUN existentes, para mantener la distribución y el paralelismo.

Segmentación de LVM

La segmentación de LVM hace referencia a distribuir datos entre varias LUN. El resultado es una mejora espectacular del rendimiento en muchas bases de datos.

Antes de la era de las unidades flash, se utilizaba la segmentación para ayudar a superar las limitaciones de rendimiento de las unidades giratorias. Por ejemplo, si un sistema operativo necesita realizar una operación de lectura de 1MB KB, para leer que 1MB TB de datos de una sola unidad se requeriría buscar y leer muchos cabezales de unidad ya que 1MB se transfiere lentamente. Si esos 1MB TB de datos se segmentaron en 8 LUN, el sistema operativo podría emitir ocho operaciones de lectura de 128K KB en paralelo y reducir el tiempo necesario para realizar la transferencia de 1MB GB.

La creación de bandas con unidades giratorias era más difícil porque el patrón de E/S debía conocerse de antemano. Si la distribución en franjas no se ajusta correctamente a los patrones de E/S reales, las configuraciones en franjas podrían dañar el rendimiento. Con las bases de datos Oracle, y especialmente con configuraciones de almacenamiento all-flash, la creación de bandas es mucho más fácil de configurar y se ha demostrado que mejora drásticamente el rendimiento.

Los gestores de volúmenes lógicos como Oracle ASM segmentan por defecto, pero el LVM del sistema operativo nativo no lo hacen. Algunos de ellos unen varias LUN como un dispositivo concatenado, lo que da como resultado archivos de datos que existen en un único dispositivo LUN. Esto provoca puntos calientes. Otras implementaciones de LVM toman por defecto extensiones distribuidas. Esto es similar a la segmentación, pero es más grueso. Las LUN del grupo de volúmenes se dividen en partes grandes, denominadas extensiones y normalmente se miden en muchos megabytes, y los volúmenes lógicos se distribuyen por esas extensiones. El resultado es que las operaciones de I/O aleatorias en un archivo se deben distribuir bien entre las LUN, pero las operaciones de I/O secuenciales no son tan eficientes como podrían.

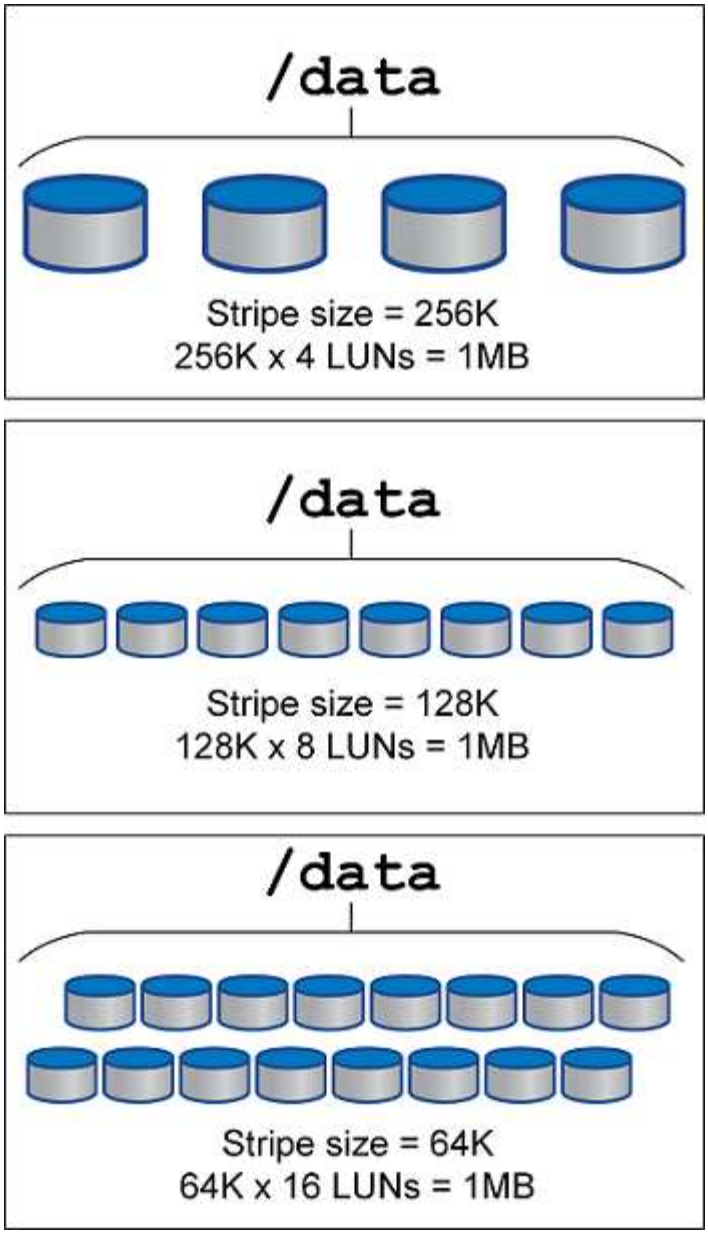
La I/O de aplicaciones con rendimiento intensivo casi siempre es una (a) en unidades del tamaño de bloque básico o (b) un megabyte.

El principal objetivo de una configuración seccionada es garantizar que la I/O de archivo único se pueda realizar como una unidad única y que las I/O de varios bloques, que deben tener un tamaño de 1MB TB, se puedan paralelizar de manera uniforme entre todas las LUN del volumen seccionado. Esto significa que el tamaño de franja no debe ser menor que el tamaño del bloque de la base de datos y el tamaño de franja multiplicado por el número de LUN debe ser 1MB.

Práctica recomendada para la creación de bandas LVM con bases de datos Oracle:

- Tamaño de franja ≥ tamaño de bloque de base de datos.
- Tamaño de franja * número de LUN ≈ 1 MB para un paralelismo óptimo.
- Utilice varios LUN por grupo de discos ASM para maximizar el rendimiento y evitar puntos críticos.

En la siguiente figura, se muestran tres opciones posibles para el ajuste del tamaño de la franja y el ancho. Se selecciona el número de LUN para satisfacer los requisitos de rendimiento tal como se han descrito anteriormente, pero en todos los casos los datos totales de una sola franja es 1MB.



NVFAIL

NVFAIL es una característica de ONTAP que garantiza la integridad de los datos durante escenarios de conmutación por error catastróficos.

Esta funcionalidad todavía se puede aplicar en los sistemas ASA r2, aunque ASA r2 utiliza una arquitectura SAN simplificada (SAZ y unidades de almacenamiento en lugar de volúmenes).

Las bases de datos son vulnerables a la corrupción durante eventos de conmutación por error de almacenamiento porque mantienen grandes cachés internos. Si un evento catastrófico requiere forzar una conmutación por error de ONTAP , independientemente del estado de la configuración general, el resultado es que los cambios previamente reconocidos pueden descartarse de manera efectiva. El contenido de la matriz de almacenamiento retrocede en el tiempo y el estado de la memoria caché de la base de datos ya no refleja el estado de los datos en el disco. Esta inconsistencia da como resultado corrupción de datos.

El almacenamiento en caché puede ocurrir en la capa de aplicación o de servidor. Por ejemplo, una configuración de Oracle Real Application Cluster (RAC) con servidores activos tanto en un sitio principal como

en uno remoto almacena en caché los datos dentro de Oracle SGA. Una operación de conmutación por error forzada que resultara en la pérdida de datos pondría la base de datos en riesgo de corrupción porque los bloques almacenados en el SGA podrían no coincidir con los bloques en el disco.

Un uso menos obvio del almacenamiento en caché es en la capa del sistema de archivos del sistema operativo. Un sistema de archivos agrupado basado en LUN ubicados en el sitio principal podría montarse en servidores en el sitio remoto y, una vez más, los datos podrían almacenarse en caché. Una falla de la NVRAM o una toma de control forzada en estas situaciones podría provocar una corrupción del sistema de archivos.

ONTAP protege las bases de datos y los sistemas operativos contra este escenario utilizando NVFAIL y sus configuraciones asociadas, que indican al host que invalide los datos en caché y vuelva a montar los sistemas de archivos afectados después de la conmutación por error. Este mecanismo se aplica a los LUN y espacios de nombres de ASA r2 tal como lo hace en AFF/ FAS.

Consideraciones clave de ASA r2:

- NVFAIL opera en el nivel LUN (unidad de almacenamiento), no en el nivel SAZ.
- Para las bases de datos Oracle, NVFAIL debe estar habilitado en todos los LUN que alojan componentes críticos (archivos de datos, registros de rehacer, archivos de control).
- MetroCluster no es compatible con ASA r2, por lo que NVFAIL se aplica principalmente a escenarios de conmutación por error de HA local.
- NFS no es compatible con ASA r2, por lo que las consideraciones de NVFAIL se aplican solo a cargas de trabajo basadas en SAN (FC/iSCSI/NVMe).

Utilidad de recuperación de ASM (ASRU)

ONTAP en ASA r2 elimina de manera eficiente los bloques en cero escritos en una LUN (unidad de almacenamiento) cuando la compresión en línea está habilitada. Las utilidades como Oracle ASM Reclamation Utility (ASRU) funcionan escribiendo ceros en las extensiones ASM no utilizadas.

Esto permite que los administradores de bases de datos recuperen espacio en la matriz de almacenamiento después de eliminar los datos. ONTAP intercepta los ceros y desasigna el espacio del LUN. El proceso de recuperación es extremadamente rápido porque no se escriben datos reales dentro del sistema de almacenamiento.

Desde el punto de vista de la base de datos, el grupo de discos de ASM contiene ceros, y leer esas regiones de las LUN daría como resultado un flujo de ceros, pero ONTAP no almacena los ceros en las unidades. En su lugar, se realizan cambios sencillos en los metadatos que marcan internamente las regiones en cero de la LUN como vacías de datos.

Por motivos similares, las pruebas de rendimiento que involucran datos puestos a cero no son válidas porque en realidad los bloques de ceros no se procesan como escrituras en la cabina de almacenamiento.

Consideraciones clave de ASRU con ASA r2 ONTAP:

- Funciona de la misma manera que AFF/ FAS para cargas de trabajo SAN porque ASA r2 es solo de bloque.
- Se aplica a LUN y espacios de nombres NVMe aprovisionados dentro de SAZ.
- No existen volúmenes FlexVol , pero el comportamiento de recuperación de bloque cero es idéntico.

Al utilizar ASRU, asegúrese de que todos los parches recomendados por Oracle están instalados.

Virtualización

La virtualización de bases de datos con VMware, Oracle OLVM o KVM es una opción cada vez más común para los clientes de NetApp que eligieron la virtualización incluso para las bases de datos más importantes.

Compatibilidad

Existen muchos malentendidos acerca de las normativas de soporte de Oracle para la virtualización, especialmente para los productos VMware. No es raro escuchar que Oracle no admite la virtualización. Esta noción es incorrecta y conduce a la pérdida de oportunidades para beneficiarse de la virtualización. El ID de documento de Oracle 249212,1 analiza los requisitos reales y los clientes rara vez consideran que son una preocupación.

Si se produce un problema en un servidor virtualizado y dicho problema es desconocido previamente para los Servicios de Soporte Oracle, es posible que se solicite al cliente que reproduzca el problema en el hardware físico. Es posible que un cliente de Oracle que ejecuta una versión de borde de sangrado de un producto no desee utilizar la virtualización debido a la posibilidad de problemas de compatibilidad, pero esta situación no ha sido un mundo real para los clientes de virtualización que utilizan versiones de productos de Oracle generalmente disponibles.

Presentación de almacenamiento

Los clientes que están considerando la virtualización de sus bases de datos deben basar sus decisiones sobre almacenamiento en sus necesidades empresariales. Aunque esta es una afirmación generalmente verdadera para todas las decisiones DE TI, es especialmente importante para los proyectos de bases de datos, porque el tamaño y el alcance de los requisitos varían considerablemente.

Existen tres opciones básicas para la presentación del almacenamiento:

- LUN virtualizados en almacenes de datos de hipervisores
- LUN iSCSI gestionadas por el iniciador iSCSI en la máquina virtual, no el hipervisor
- Sistemas de archivos NFS montados por la máquina virtual (no desde un almacén de datos basado en NFS)
- Asignaciones directas de dispositivos. Los clientes no se ven favorecidos por los RDM de VMware, pero los dispositivos físicos siguen siendo a menudo asignados de forma similar directamente con la virtualización de KVM y OLVM.

Rendimiento

El método de presentar almacenamiento a un invitado virtualizado no suele afectar al rendimiento. Todos los SO host, los controladores de red virtualizados y las implementaciones de almacenes de datos de hipervisor están muy optimizados y, por lo general, pueden consumir todo el ancho de banda de red FC o IP disponible entre el hipervisor y el sistema de almacenamiento, siempre que se sigan las prácticas recomendadas básicas. En algunos casos, obtener un rendimiento óptimo puede ser ligeramente más sencillo usando un método de presentación de almacenamiento en comparación con otro, pero el resultado final debería ser comparable.

Gran capacidad de administración

El factor clave para decidir cómo presentar el almacenamiento a un invitado virtualizado es la capacidad de gestión. No hay un método correcto o incorrecto. El mejor enfoque depende de las necesidades operativas, las habilidades y las preferencias de TI.

Los factores a considerar incluyen:

- **Transparencia.** Cuando una VM administra sus sistemas de archivos, es más fácil para un administrador de bases de datos o un administrador del sistema identificar el origen de los sistemas de archivos para sus datos. No se accede a los sistemas de archivos y LUN de manera diferente a con un servidor físico.
- **Consistencia.** Cuando una VM es propietaria de sus sistemas de archivos, el uso o no uso de una capa de hipervisor afecta a la capacidad de gestión. Los mismos procedimientos para aprovisionamiento, supervisión, protección de datos, etc. se pueden utilizar en todo el conjunto, incluidos los entornos virtualizados y no virtualizados.

  Por otro lado, en un centro de datos virtualizado al 100% de lo contrario, puede que sea preferible también utilizar el almacenamiento basado en almacenes de datos en toda la huella en la misma razón mencionada anteriormente: Consistencia: La capacidad de usar los mismos procedimientos para aprovisionamiento, protección, montorización y protección de datos.

- **Estabilidad y resolución de problemas.** Cuando una VM es propietaria de sus sistemas de archivos, ofrecer un rendimiento bueno y estable y solucionar problemas son más simples porque toda la pila de almacenamiento está presente en la VM. El único rol del hipervisor es transportar tramas FC o IP. Cuando se incluye un almacén de datos en una configuración, esto complica la configuración introduciendo otro conjunto de tiempos de espera, parámetros, archivos de registro y posibles errores.
- **Portabilidad.** Cuando una VM posee sus sistemas de archivos, el proceso de mover un entorno de Oracle se vuelve mucho más sencillo. Los sistemas de archivos se pueden mover fácilmente entre huéspedes virtualizados y no virtualizados.
- * Bloqueo del proveedor.* Después de colocar los datos en un almacén de datos, usar un hipervisor diferente o extraer los datos del entorno virtualizado se vuelve completamente difícil.
- **Activación de instantáneas.** Los procedimientos de respaldo tradicionales en un entorno virtualizado pueden convertirse en un problema debido al ancho de banda relativamente limitado. Por ejemplo, un tronco 10GbE de cuatro puertos podría ser suficiente para soportar las necesidades de rendimiento diarias de muchas bases de datos virtualizadas, pero tal tronco sería insuficiente para realizar copias de seguridad con RMAN u otros productos de copia de seguridad que requieran transmitir una copia de tamaño completo de los datos. El resultado es que un entorno virtualizado cada vez más consolidado debe realizar backups a través de snapshots de almacenamiento. Esto evita la necesidad de sobrecargar la configuración del hipervisor únicamente para admitir los requisitos de ancho de banda y CPU de la ventana de backup.

  El uso de sistemas de archivos propiedad de invitados a veces facilita el uso de backups y restauraciones basados en copias Snapshot, ya que los objetos de almacenamiento que necesitan protección pueden dirigirse con mayor facilidad. Sin embargo, cada vez es más grande la cantidad de productos de protección de datos de virtualización que se integran bien con los almacenes de datos y las copias Snapshot. La estrategia de backup debe consistir completamente antes de tomar una decisión sobre cómo presentar el almacenamiento a un host virtualizado.

Controladores paravirtualizados

Para un rendimiento óptimo, el uso de controladores de red paravirtualizados es fundamental. Cuando se utiliza un almacén de datos, se requiere un controlador SCSI paravirtualizado. Un controlador de dispositivo

paravirtualizado permite a un invitado integrarse más profundamente en el hipervisor, en lugar de un controlador emulado en el que el hipervisor pasa más tiempo de CPU imitando el comportamiento del hardware físico.

RAM de sobrecompromiso

Sobrecomprometer RAM significa configurar más RAM virtualizada en varios hosts de la que existe en el hardware físico. Si lo hace, se pueden producir problemas de rendimiento inesperados. Al virtualizar una base de datos, el hipervisor no debe intercambiar los bloques subyacentes del SGA de Oracle en el almacenamiento. Si lo hace, los resultados de rendimiento son muy inestables.

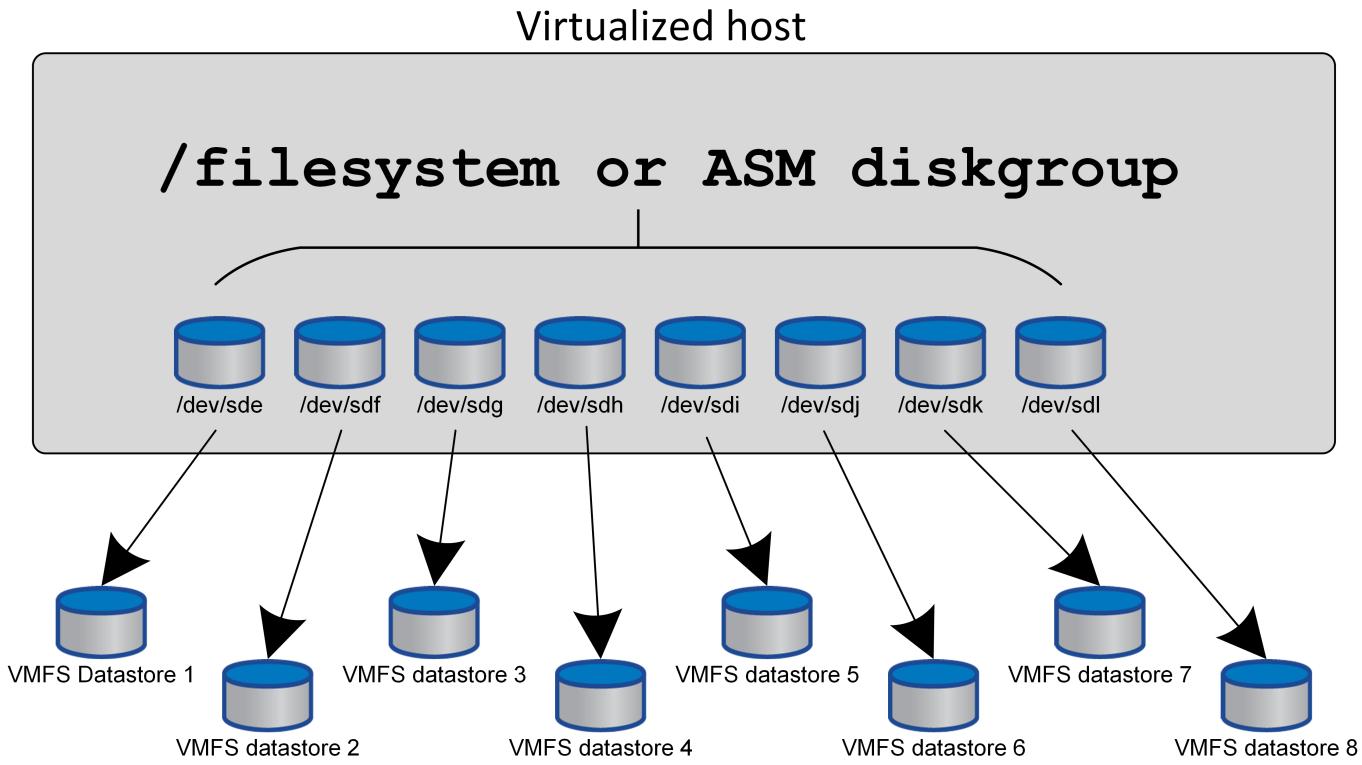
Segmentación de almacenes de datos

Cuando se usan bases de datos con almacenes de datos, hay un factor crucial que debe tenerse en cuenta con respecto al rendimiento: La segmentación.

Las tecnologías de almacenes de datos como VMFS pueden abarcar varios LUN, pero no son dispositivos segmentados. Las LUN se concatenan. El resultado final pueden ser puntos de sobrecarga de la LUN. Por ejemplo, una base de datos de Oracle típica puede tener un grupo de discos ASM de 8 LUN. Se pueden aprovisionar los 8 LUN virtualizados en un almacén de datos VMFS de 8 LUN, pero no hay garantía de cuáles LUN residirán los datos. La configuración resultante podría ser todos los 8 LUN virtualizados que ocupen una única LUN dentro del almacén de datos VMFS. Esto se convierte en un cuello de botella en el rendimiento.

La segmentación suele ser necesaria. Con algunos hipervisores, incluido KVM, es posible crear un almacén de datos con la segmentación de LVM, como se describe "aquí". Con VMware, la arquitectura parece un poco diferente. Cada LUN virtualizado debe colocarse en un almacén de datos VMFS diferente.

Por ejemplo:



El impulsor principal de este enfoque no es ONTAP, sino que se debe a una limitación inherente al número de

operaciones que una sola máquina virtual o LUN de hipervisor puede prestar servicio en paralelo. Por lo general, una sola LUN de ONTAP puede admitir muchas más IOPS de las que puede solicitar un host. El límite de rendimiento de una LUN es casi universal debido al SO del host. Como resultado, la mayoría de las bases de datos necesitan entre 4 y 8 LUN para satisfacer sus necesidades de rendimiento.

Las arquitecturas de VMware deben planificar sus arquitecturas con cuidado para asegurarse de que no se encuentren los máximos de almacén de datos o ruta de LUN con este enfoque. Además, no es necesario disponer de un conjunto único de almacenes de datos VMFS para cada base de datos. La principal necesidad es asegurarse de que cada host tenga un conjunto limpio de 4-8 rutas de I/O desde las LUN virtualizadas hasta las LUN de back-end del sistema de almacenamiento propiamente dicho. En raras ocasiones, incluso más almacenes de datos pueden ser útiles para las demandas de rendimiento realmente extremas, pero 4-8 LUN suelen ser suficientes para el 95% de todas las bases de datos. Un solo volumen ONTAP que contiene 8 LUN puede admitir hasta 250.000 IOPS de bloques de Oracle aleatorias con una configuración típica de SO/ONTAP/red.

Organización en niveles

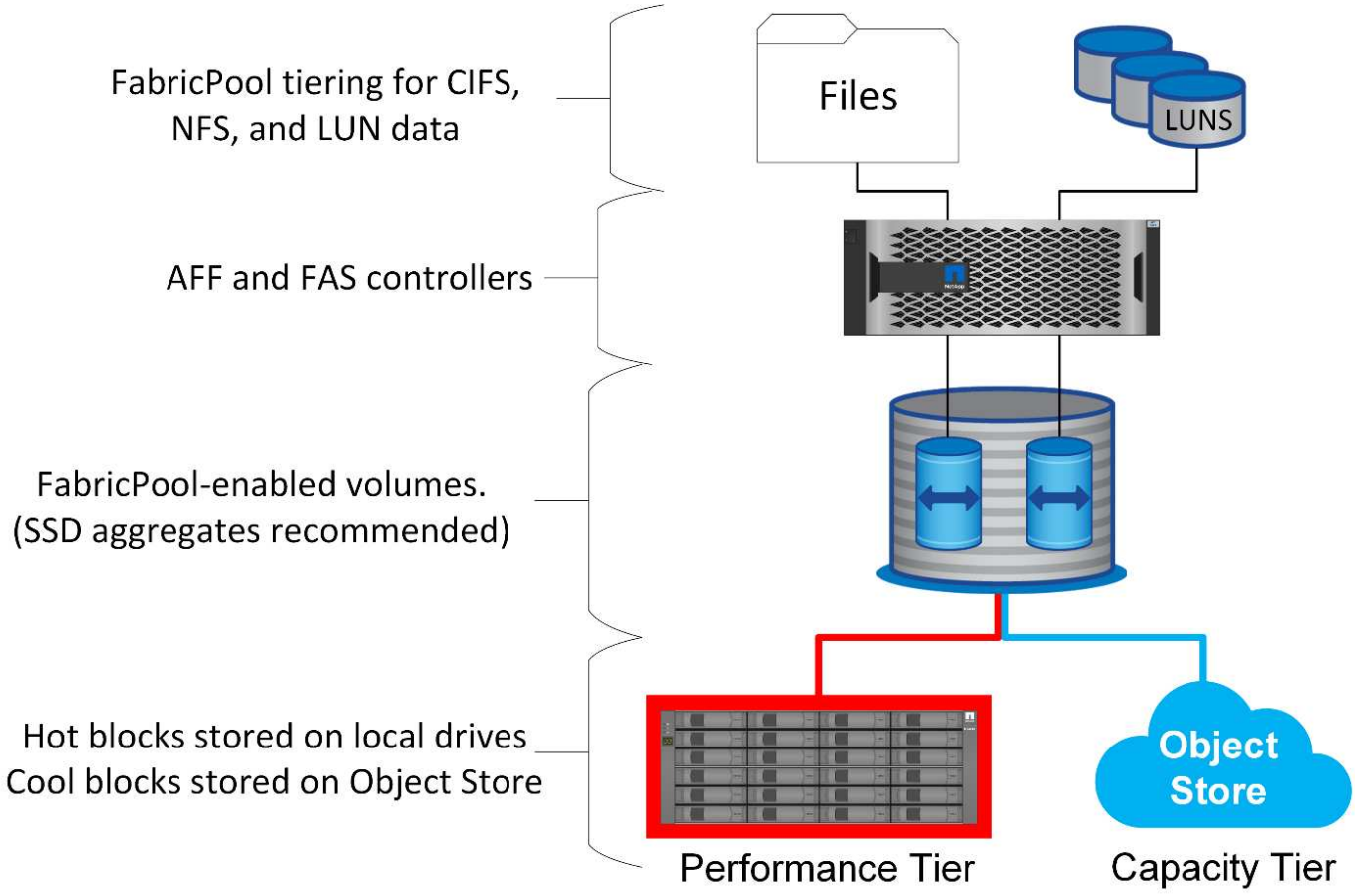
Descripción general

Comprender cómo afecta el almacenamiento por niveles FabricPool a Oracle y otras bases de datos requiere comprender la arquitectura de FabricPool de bajo nivel.

Arquitectura

FabricPool es una tecnología de organización en niveles que clasifica los bloques como activos o inactivos y los coloca en el nivel de almacenamiento más adecuado. El nivel de rendimiento con mayor frecuencia se encuentra en el almacenamiento SSD y aloja los bloques de datos activos. El nivel de capacidad está ubicado en un almacén de objetos y aloja los bloques de datos inactivos. La compatibilidad de almacenamiento de objetos incluye NetApp StorageGRID, ONTAP S3, almacenamiento Microsoft Azure Blob, servicio de almacenamiento de objetos en el cloud de Alibaba, almacenamiento de objetos de IBM Cloud, almacenamiento de Google Cloud y Amazon AWS S3.

Existen varias políticas de organización en niveles disponibles que controlan la clasificación de los bloques como activos o inactivos, y las políticas se pueden establecer por volumen y modificar según sea necesario. Solo se mueven los bloques de datos entre los niveles de rendimiento y capacidad. Los metadatos que definen la LUN y la estructura del sistema de archivos siempre permanecen en el nivel de rendimiento. Como resultado, la gestión se centraliza en ONTAP. Los archivos y los LUN no aparecen diferentes de los datos almacenados en cualquier otra configuración de ONTAP. La controladora NetApp AFF o FAS aplica las políticas definidas para mover datos al nivel adecuado.



Proveedores de almacenes de objetos

Los protocolos de almacenamiento de objetos utilizan solicitudes HTTP o HTTPS sencillas para almacenar grandes cantidades de objetos de datos. El acceso al almacenamiento de objetos debe ser fiable, porque el acceso a los datos desde ONTAP depende de atender solicitudes rápidamente. Entre las opciones se incluyen las opciones de acceso estándar y poco frecuente de Amazon S3, y Microsoft Azure Hot and Cool Blob Storage, IBM Cloud y Google Cloud. No se admiten opciones de archivado como Amazon Glacier y Amazon Archive porque el tiempo necesario para recuperar los datos puede superar las tolerancias de las aplicaciones y los sistemas operativos del host.

También se ofrece compatibilidad con NetApp StorageGRID y es una solución empresarial óptima. Es un sistema de almacenamiento de objetos de alto rendimiento, escalable y altamente seguro que puede proporcionar redundancia geográfica para los datos de FabricPool, así como otras aplicaciones de almacenamiento de objetos que tienen cada vez más probabilidades de formar parte de entornos de aplicaciones empresariales.

StorageGRID también puede reducir los costes al evitar los cargos por salida que imponen muchos proveedores de cloud público por leer los datos de sus servicios.

Los datos y metadatos

Tenga en cuenta que el término «datos» aquí se aplica a los bloques de datos reales, no a los metadatos. Solo los bloques de datos se organizan en niveles, mientras que los metadatos permanecen en el nivel de rendimiento. Además, el estado de un bloque como activo o inactivo solo se ve afectado por la lectura del bloque de datos real. La simple lectura del nombre, la marca de tiempo o los metadatos de propiedad de un archivo no afecta a la ubicación de los bloques de datos subyacentes.

Completos

Aunque FabricPool puede reducir significativamente el espacio físico de almacenamiento, no es por sí misma una solución de backup. Los metadatos de NetApp WAFL siempre permanecen en el nivel de rendimiento. Si un desastre catastrófico destruye el nivel de rendimiento, no se puede crear un nuevo entorno con los datos del nivel de capacidad porque no contiene metadatos de WAFL.

Sin embargo, FabricPool puede formar parte de una estrategia de backup. Por ejemplo, FabricPool se puede configurar con la tecnología de replicación SnapMirror de NetApp. Cada mitad del reflejo puede tener su propia conexión con un destino de almacenamiento de objetos. El resultado es dos copias independientes de los datos. La copia primaria consiste en los bloques del nivel de rendimiento y los bloques asociados del nivel de capacidad, y la réplica es un segundo conjunto de bloques de rendimiento y capacidad.

Políticas de organización en niveles

Políticas de organización en niveles

ONTAP tiene disponibles cuatro políticas que controlan cómo los datos de Oracle en el nivel de rendimiento se convierten en candidatos para reubicar al nivel de capacidad.

Solo Snapshot

La `snapshot-only tiering-policy` se aplica sólo a los bloques que no se comparten con el sistema de archivos activo. Básicamente, provoca la organización en niveles de los backups de las bases de datos. Los bloques se convierten en candidatos para organizar por niveles después de que se crea una copia Snapshot y se sobrescribe el bloque, lo que genera un bloque que solo existe dentro de la copia Snapshot. El retraso antes de a. `snapshot-only` el bloque se considera frío y está controlado por el `tiering-minimum-cooling-days` configuración para el volumen. El intervalo a partir de ONTAP 9,8 es de 2 a 183 días.

Muchos conjuntos de datos tienen tasas de cambio bajas, lo que resulta en un ahorro mínimo de esta política. Por ejemplo, una base de datos típica observada en ONTAP tiene una tasa de cambio inferior al 5% a la semana. Los archive logs de la base de datos pueden ocupar mucho espacio, pero normalmente continúan existiendo en el sistema de archivos activo y, por lo tanto, no serían candidatos para la organización en niveles bajo esta política.

Automático

La `auto` la política de organización en niveles amplía la clasificación por niveles tanto a bloques específicos de snapshots como a bloques del sistema de archivos activo. El retardo antes de que un bloque se considere frío es controlado por el `tiering-minimum-cooling-days` configuración para el volumen. El intervalo a partir de ONTAP 9,8 es de 2 a 183 días.

Este método permite opciones de organización en niveles que no están disponibles con el `snapshot-only` política. Por ejemplo, una política de protección de datos puede requerir 90 días de ciertos archivos de registro para ser retenidos. Si se establece un período de enfriamiento de 3 días, los archivos de registro anteriores a 3 días se almacenarán en niveles desde la capa de rendimiento. Esta acción libera espacio considerable en el nivel de rendimiento a la vez que le permite ver y gestionar los 90 días completos de datos.

Ninguno

La `none` la política de organización en niveles evita que cualquier bloque adicional se organice en niveles desde la capa de almacenamiento, pero todos los datos que permanezcan en el nivel de capacidad permanecen en el nivel de capacidad hasta que se leen. Si a continuación se lee el bloque, se retira y se coloca en el nivel de rendimiento.

El motivo principal para utilizar el `none` la política de organización en niveles es para evitar que los bloques se organicen en niveles, pero podría resultar útil cambiar las políticas con el tiempo. Por ejemplo, pongamos por caso que un conjunto de datos concreto se organiza ampliamente en niveles en la capa de capacidad, pero surge una necesidad inesperada de funcionalidades de rendimiento completas. La política se puede cambiar para evitar cualquier organización en niveles adicional y para confirmar que los bloques que se lean a medida que los aumentos de I/O permanecen en el nivel de rendimiento.

Todo

La `all` la política de organización en niveles reemplaza el `backup` Normativa a partir de ONTAP 9,6. La `backup` Política aplicada solo a los volúmenes de protección de datos, lo que significa un destino de SnapMirror o NetApp SnapVault. La `all` la política funciona de la misma manera, pero no se limita a los volúmenes de protección de datos.

Con esta política, los bloques se consideran inmediatamente inactivos y elegibles para organizarse en niveles en la capa de capacidad de inmediato.

Esta política resulta especialmente adecuada para backups a largo plazo. También se puede utilizar como una forma de gestión de almacenamiento jerárquico (HSM). Anteriormente, se utilizaba HSM para organizar en niveles los bloques de datos de un archivo en cinta y, al mismo tiempo, mantener el propio archivo visible en el sistema de archivos. Un volumen FabricPool con el `all` la política le permite almacenar archivos en un nivel visible y gestionable pero consume prácticamente ningún espacio en el nivel de almacenamiento local.

Políticas de recuperación

Las políticas de organización en niveles controlan qué bloques de la base de datos de Oracle se organizan en niveles desde el nivel de rendimiento al nivel de capacidad. Las políticas de recuperación controlan lo que sucede cuando se lee un bloque que se ha organizado en niveles.

Predeterminado

Inicialmente, todos los volúmenes FabricPool se establecen en `default`, que significa que el comportamiento está controlado por la `política de recuperación de nubes. `El comportamiento exacto depende de la política de organización en niveles utilizada.

- `auto`- solo recuperar datos de lectura aleatoria
- `snapshot-only`- recuperar todos los datos de lectura secuencial o aleatoria
- `none`- recuperar todos los datos de lectura secuencial o aleatoria
- `all`- no recuperar datos del nivel de capacidad

En lectura

Ajuste `cloud-retrieval-policy` en la lectura sobrescribe el comportamiento predeterminado, de modo que la lectura de cualquier dato por niveles provoca que esos datos se devuelvan al nivel de rendimiento.

Por ejemplo, es posible que un volumen se haya usado ligeramente durante mucho tiempo en `auto` la política de organización en niveles y la mayoría de los bloques están ahora organizados en niveles.

Si un cambio inesperado en las necesidades empresariales requirió que algunos de los datos se escanearan repetidamente para preparar un determinado informe, puede ser conveniente cambiar el `cloud-retrieval-`

`policy` para `on-read` para garantizar que todos los datos que se leen se devuelven al nivel de rendimiento, incluidos datos de lectura secuencial y aleatoria. Esto mejoraría el rendimiento de I/O secuenciales en el volumen.

Promocione

El comportamiento de la política de promoción depende de la política de organización en niveles. Si la política de organización en niveles es `auto`, y, a continuación, ajuste el `cloud-retrieval-policy `to `promote` devuelve todos los bloques del nivel de capacidad en el siguiente análisis de organización en niveles.

Si la política de organización en niveles es `snapshot-only`, entonces, los únicos bloques que se devuelven son los bloques asociados al sistema de archivos activo. Normalmente, esto no tendría ningún efecto porque los únicos bloques organizados en niveles en `snapshot-only` la política sería bloques asociados exclusivamente a las instantáneas. No habría bloques por niveles en el sistema de archivos activo.

Sin embargo, si un SnapRestore de volumen o una operación de clonado de archivos se restauraron los datos de un volumen desde una copia Snapshot, es posible que algunos de los bloques organizados en niveles debido a que solo estaban asociados a snapshots ahora sean requeridos por el sistema de archivos activo. Puede ser conveniente cambiar temporalmente el `cloud-retrieval-policy` política a. `promote` para recuperar rápidamente todos los bloques necesarios localmente.

Nunca

No recupere bloques del nivel de capacidad.

Estrategias de organización en niveles

Organización en niveles de archivos completa

Aunque la organización en niveles de FabricPool opera a nivel de bloques, en algunos casos se puede utilizar para la organización en niveles de archivos.

Muchas aplicaciones están organizadas por fecha, y por lo general es menos probable que se acceda a estos datos a medida que envejecen. Por ejemplo, un banco puede tener un repositorio de archivos PDF que contenga cinco años de extractos de clientes, pero sólo están activos los últimos meses. FabricPool se puede usar para reubicar archivos de datos más antiguos en el nivel de capacidad. Un período de enfriamiento de 14 días garantizaría que los 14 días más recientes de archivos PDF permanezcan en el nivel de rendimiento. Además, los archivos que se leen al menos cada 14 días permanecerán activos y, por consiguiente, permanecerán en el nivel de rendimiento.

Normativas

Para implementar un método de organización en niveles basado en archivos, debe tener archivos que se escriban y no se modifiquen posteriormente. La `tiering-minimum-cooling-days` la política debe establecerse lo suficientemente alta para que los archivos que pueda necesitar permanezcan en el nivel de rendimiento. Por ejemplo, un conjunto de datos para los que se requieren los 60 días de datos más recientes y un rendimiento óptimo garantiza configurar el `tiering-minimum-cooling-days` hasta 60. También se pueden obtener resultados similares en función de los patrones de acceso a archivos. Por ejemplo, si se requieren los últimos 90 días de datos y la aplicación accede a ese intervalo de 90 días, los datos permanecerán en el nivel de rendimiento. Mediante la configuración de `tiering-minimum-cooling-days` en el periodo 2, se obtiene una organización en niveles inmediata después de que los datos se vuelven menos activos.

La `auto` se requiere una política para impulsar la organización en niveles de estos bloques porque solo el `auto` la política afecta a los bloques que están en el sistema de archivos activo.

> Cualquier tipo de acceso a los datos restablece los datos del mapa de calor. La detección de virus, la indexación e incluso la actividad de backup que lee los archivos de origen evita la segmentación, ya que es necesario `tiering-minimum-cooling-days` nunca se ha alcanzado el umbral.

Organización parcial en niveles de archivos

Dado que FabricPool funciona a nivel de bloque, los archivos que están sujetos a cambios se pueden organizar parcialmente en niveles en el almacenamiento de objetos y, al mismo tiempo, permanecen parcialmente en el nivel de rendimiento.

Esto es común con las bases de datos. Las bases de datos que se sabe que contienen bloques inactivos también son candidatas para la organización en niveles de FabricPool. Por ejemplo, una base de datos de gestión de cadena de suministro puede contener información histórica que debe estar disponible si es necesario, pero que no se puede acceder durante las operaciones normales. FabricPool se puede utilizar para reubicar selectivamente los bloques inactivos.

Por ejemplo, los archivos de datos que se ejecutan en un volumen FabricPool con a. `tiering-minimum-cooling-days` periodo de 90 días: conserva los bloques a los que se ha accedido en los 90 días anteriores en el nivel de rendimiento. Sin embargo, todo lo que no se acceda durante 90 días se reubica al nivel de capacidad. En otros casos, la actividad normal de la aplicación conserva los bloques correctos en el nivel correcto. Por ejemplo, si una base de datos se utiliza normalmente para procesar los 60 días anteriores de datos de forma regular, es mucho menor `tiering-minimum-cooling-days` el período se puede establecer porque la actividad natural de la aplicación garantiza que los bloques no se reubiquen antes de tiempo.

> La `auto` la política debe utilizarse con cuidado con las bases de datos. Muchas bases de datos tienen actividades periódicas como el proceso de final del trimestre o las operaciones de reindexación. Si el período de estas operaciones es mayor que el `tiering-minimum-cooling-days` se pueden producir problemas de rendimiento. Por ejemplo, si el procesamiento a final de trimestre requiere 1TB TB de datos que de otro modo no se han modificado, esos datos podrían estar presentes ahora en el nivel de capacidad. Las lecturas del nivel de capacidad a menudo son extremadamente rápidas y pueden no causar problemas de rendimiento, pero los resultados exactos dependerán de la configuración del almacén de objetos.

Normativas

La `tiering-minimum-cooling-days` la política debe establecerse lo suficientemente alta para conservar los archivos que pueden ser necesarios en el nivel de rendimiento. Por ejemplo, una base de datos en la que los 60 días de datos más recientes podrían ser necesarios con un rendimiento óptimo justificaría establecer el `tiering-minimum-cooling-days` periodo hasta 60 días. También se podrían lograr resultados similares en función de los patrones de acceso de los archivos. Por ejemplo, si se requieren los 90 días de datos más recientes y la aplicación accede a ese intervalo de 90 días de datos, los datos permanecerán en el nivel de rendimiento. Ajuste de `tiering-minimum-cooling-days` un periodo de hasta 2 días clasificaría los datos en niveles inmediatamente después de que los datos se vuelvan menos activos.

La `auto` se requiere una política para impulsar la organización en niveles de estos bloques porque solo el `auto` la política afecta a los bloques que están en el sistema de archivos activo.

Cualquier tipo de acceso a los datos restablece los datos del mapa de calor. Por lo tanto, las exploraciones de tablas completas de la base de datos e incluso la actividad de copia de seguridad que lee los archivos de origen impiden la organización en niveles porque es necesario `tiering-minimum-cooling-days` nunca se ha alcanzado el umbral.

Organización en niveles de archive log

Quizás el uso más importante de FabricPool sea mejorar la eficiencia de los datos fríos conocidos, como los registros de transacciones de base de datos.

La mayoría de las bases de datos relacionales funcionan en modo de archivado de registros de transacciones para ofrecer una recuperación puntual. Los cambios en las bases de datos se confirman registrando los cambios en los registros de transacciones y el registro de transacciones se conserva sin sobrescribirse. El resultado, puede ser un requisito para conservar un enorme volumen de registros de transacciones archivados. Existen ejemplos similares con muchos otros flujos de trabajo de aplicaciones que generan datos que deben conservarse, pero es muy poco probable que se acceda jamás.

FabricPool resuelve estos problemas al ofrecer una única solución con organización en niveles integrada. Los archivos se almacenan y siguen siendo accesibles en su ubicación habitual, pero prácticamente no ocupan espacio en la matriz primaria.

Normativas

Utilice un `tiering-minimum-cooling-days` la política de unos días provoca una retención de bloques en los archivos creados recientemente (que son los archivos con mayor probabilidad de que sean necesarios a corto plazo) en el nivel de rendimiento. Los bloques de datos de los archivos antiguos se mueven al nivel de capacidad.

La `auto` aplica la clasificación por niveles de avisos cuando se alcanza el umbral de enfriamiento independientemente de si los registros se han suprimido o siguen existiendo en el sistema de archivos primario. También se simplifica la gestión, almacenar todos los registros potencialmente necesarios en una sola ubicación del sistema de archivos activo. No hay razón para buscar a través de instantáneas para localizar un archivo que necesita ser restaurado.

Algunas aplicaciones, como Microsoft SQL Server, truncan los archivos de registro de transacciones durante las operaciones de backup de modo que los registros ya no estén en el sistema de archivos activo. Se puede ahorrar capacidad mediante el uso del `snapshot-only` la política de organización en niveles, solo el `auto` la política no es útil para los datos de registro porque rara vez deben enfriarse los datos de registro en el sistema de archivos activo.

Organización en niveles de Snapshot

La versión inicial de FabricPool se dirigía al caso de uso de backup. El único tipo de bloques que se podía organizar en niveles eran bloques que ya no estaban asociados a los datos del sistema de archivos activo. Por lo tanto, solo se pueden mover los bloques de datos de Snapshot al nivel de capacidad. Esta sigue siendo una de las opciones de organización en niveles más seguras cuando hay que garantizar que el rendimiento nunca se vea afectado.

Políticas: Snapshots locales

Existen dos opciones para organizar en niveles los bloques Snapshot inactivos en el nivel de capacidad. En

primer lugar, el `snapshot-only` la política solo apunta a los bloques de instantáneas. Aunque la `auto` la política incluye el `snapshot-only` bloques, también organiza en niveles bloques del sistema de archivos activo. Esto podría no ser deseable.

La `tiering-minimum-cooling-days` el valor se debe establecer en un período de tiempo que permita que los datos que pueden necesitarse durante una restauración estén disponibles en el nivel de rendimiento. Por ejemplo, la mayoría de los escenarios de restauración de una base de datos de producción crucial incluyen un punto de restauración en algún momento en los pocos días anteriores. Ajuste A `tiering-minimum-cooling-days` el valor de 3 garantizaría que cualquier restauración del archivo da como resultado un archivo que proporciona un rendimiento máximo inmediatamente. Todos los bloques de los archivos activos se encuentran presentes en un almacenamiento rápido sin necesidad de recuperarlos del nivel de capacidad.

Políticas: Snapshots replicadas

Un snapshot que se replica con SnapMirror o SnapVault solo se usa para la recuperación deberá utilizar FabricPool `all` política. Con esta política, los metadatos se replican, pero todos los bloques de datos se envían inmediatamente al nivel de capacidad, lo que genera un rendimiento máximo. La mayoría de los procesos de recuperación implican una I/O secuencial, que es inherentemente eficiente. El tiempo de recuperación del almacén de objetos se debe evaluar, pero en una arquitectura bien diseñada, no es necesario que este proceso de recuperación sea significativamente más lento que la recuperación de datos locales.

Si también se van a usar los datos replicados para la clonación, el `auto` la política es más apropiada, con un `tiering-minimum-cooling-days` valor que abarca los datos que se espera que se utilicen regularmente en un entorno de clonación. Por ejemplo, el conjunto de trabajo activo de una base de datos puede incluir datos leídos o escritos en los tres días anteriores, pero también podría incluir otros 6 meses de datos históricos. Si es así, entonces el `auto` En el destino de SnapMirror, el conjunto de trabajo está disponible en el nivel de rendimiento.

Organización en niveles de backup

Las copias de seguridad de aplicaciones tradicionales incluyen productos como Oracle Recovery Manager, que crea copias de seguridad basadas en archivos fuera de la ubicación de la base de datos original.

```
 `tiering-minimum-cooling-days` policy of a few days preserves the most
recent backups, and therefore the backups most likely to be required for
an urgent recovery situation, on the performance tier. The data blocks of
the older files are then moved to the capacity tier.
La `auto` la política más adecuada para los datos de backup. Esto
garantiza la clasificación por niveles de avisos cuando se ha alcanzado el
umbral de enfriamiento independientemente de si los archivos se han
suprimido o siguen existiendo en el sistema de archivos primario. También
simplifica la gestión almacenar todos los archivos potencialmente
necesarios en una sola ubicación del sistema de archivos activo. No hay
razón para buscar a través de instantáneas para localizar un archivo que
necesita ser restaurado.
```

La `snapshot-only` la política podría funcionar, pero esa política solo se aplica a los bloques que ya no están en el sistema de archivos activo. Por lo tanto, los archivos en un recurso compartido NFS o SMB deben

eliminarse primero para poder organizar los datos en niveles.

Esta política sería aún menos eficiente con una configuración de LUN porque la eliminación de un archivo de una LUN solo elimina las referencias de archivos de los metadatos del sistema de archivos. Los bloques reales de las LUN permanecen en su lugar hasta que se sobrescriben. Esta situación puede crear un retraso prolongado entre el momento en que se elimina un archivo y el tiempo que se sobrescriben los bloques y se convierten en candidatos para la organización en niveles. El traslado de la `snapshot-only` Bloques en el nivel de capacidad pero, en general, la gestión de datos de backup de FabricPool funciona mejor con el `auto` política.

Este enfoque ayuda a los usuarios a gestionar el espacio necesario para los backups de una forma más eficiente, pero el propio FabricPool no es una tecnología de backup. La organización en niveles de los archivos de backup en el almacén de objetos simplifica la gestión, ya que los archivos siguen visibles en el sistema de almacenamiento original, pero los bloques de datos del destino del almacén de objetos dependen del sistema de almacenamiento original. Si se pierde el volumen de origen, los datos del almacén de objetos ya no se pueden usar.

Interrupciones de acceso al almacén de objetos

La organización en niveles de un conjunto de datos con FabricPool provoca una dependencia entre la cabina de almacenamiento principal y el nivel de almacén de objetos. Hay muchas opciones de almacenamiento de objetos que ofrecen distintos niveles de disponibilidad. Es importante comprender el impacto de una posible pérdida de conectividad entre la cabina de almacenamiento primaria y el nivel de almacenamiento de objetos.

Si una I/O emitida a ONTAP requiere datos del nivel de capacidad y ONTAP no puede alcanzar el nivel de capacidad para recuperar los bloques, se agotará el tiempo de espera de las I/O finalmente. El efecto de este tiempo de espera depende del protocolo utilizado. En un entorno NFS, ONTAP responde con una respuesta EJUKEBOX o EDELAY, dependiendo del protocolo. Algunos sistemas operativos anteriores pueden interpretarlo como un error, pero los sistemas operativos actuales y los niveles de parches actuales del cliente Oracle Direct NFS lo tratan como un error recuperable y siguen esperando a que se complete la E/S.

Un tiempo de espera menor se aplica a los entornos SAN. Si se requiere un bloque en el entorno de almacén de objetos y permanece inaccesible durante dos minutos, se devuelve un error de lectura al host. El volumen ONTAP y los LUN permanecen en línea, pero el SO del host puede marcar el sistema de archivos como está en estado de error.

Problemas de conectividad del almacenamiento de objetos `snapshot-only` la política es menos problemática, ya que únicamente los datos de backup están organizados en niveles. Los problemas de comunicación ralentizarían la recuperación de datos, pero de otro modo no afectarían a los datos que se están utilizando activamente. La `auto` y.. `all` Las políticas permiten la clasificación por niveles de los datos inactivos de la LUN activa, lo que significa que un error durante la recuperación de datos del almacén de objetos puede afectar a la disponibilidad de la base de datos. Una implementación de SAN con estas políticas solo debe utilizarse con almacenamiento de objetos de clase empresarial y conexiones de red diseñadas para obtener una alta disponibilidad. NetApp StorageGRID es la opción superior.

Protección de datos de Oracle

Protección de datos con ONTAP

NetApp sabe que los datos más críticos se encuentran en las bases de datos.

Una empresa no puede operar sin acceso a sus datos y, a veces, los datos definen el negocio. Estos datos deben protegerse; sin embargo, la protección de datos no solo garantiza un backup utilizable; se trata de realizar backups de forma rápida y fiable, además de almacenarlos de forma segura.

El otro lado de la protección de datos es la recuperación de datos. Cuando no se puede acceder a los datos, la empresa se ve afectada y puede dejar de funcionar hasta que se restauren los datos. Este proceso debe ser rápido y fiable. Por último, la mayoría de las bases de datos deben protegerse frente a desastres, lo que significa mantener una réplica de la base de datos. La réplica debe estar lo suficientemente actualizada. También debe ser rápido y sencillo hacer de la réplica una base de datos completamente operativa.

Esta documentación sustituye al informe técnico *TR-4591 publicado anteriormente: Protección de datos de Oracle: Backup, recuperación y replicación.*

Planificación

La arquitectura de protección de datos empresariales adecuada depende de los requisitos empresariales relacionados con la retención de datos, la capacidad de recuperación y la tolerancia a interrupciones durante diversos eventos.

Por ejemplo, piense en el número de aplicaciones, bases de datos y conjuntos de datos importantes. Crear una estrategia de backup para un único conjunto de datos que garantice el cumplimiento de los acuerdos de nivel de servicio típicos es bastante sencillo, ya que no hay muchos objetos que gestionar. A medida que aumenta el número de conjuntos de datos, la supervisión se hace más complicada y los administradores pueden verse forzados a invertir cada vez más tiempo en solucionar los fallos de backup. A medida que un entorno llega al cloud y escala el proveedor de servicios, se necesita un enfoque totalmente diferente.

El tamaño del conjunto de datos también afecta a la estrategia. Por ejemplo, existen muchas opciones para backup y recuperación con una base de datos 100GB porque el conjunto de datos es tan pequeño. La simple copia de los datos de los medios de backup con herramientas tradicionales suele proporcionar un objetivo de tiempo de recuperación suficiente para la recuperación. Una base de datos de 100TB suele necesitar una estrategia completamente diferente a menos que el objetivo de tiempo de recuperación permita una interrupción de varios días, en cuyo caso puede ser aceptable un procedimiento tradicional de backup y recuperación basado en copia.

Por último, existen factores fuera del propio proceso de backup y recuperación. Por ejemplo, ¿existen bases de datos que respalden actividades de producción críticas, lo que convierte la recuperación en un evento raro que solo realizan los administradores de bases de datos cualificados? Alternativamente, ¿las bases de datos forman parte de un entorno de desarrollo de gran tamaño en el que la recuperación es una ocurrencia frecuente y gestionada por un EQUIPO de TECNOLOGÍA generalista?

Planificación de objetivos de tiempo, objetivos de punto de recuperación y acuerdos de nivel de servicio

ONTAP le permite adaptar con facilidad una estrategia de protección de datos de base de datos de Oracle a sus requisitos empresariales.

Entre estos requisitos se incluyen factores como la velocidad de recuperación, la pérdida de datos máxima permitida y las necesidades de retención de backup. El plan de protección de datos también debe tener en cuenta varios requisitos normativos para la retención y restauración de datos. Por último, deben tenerse en

cuenta diferentes escenarios de recuperación de datos, que van desde la recuperación típica y previsible que se produce por errores de usuarios o aplicaciones hasta escenarios de recuperación de desastres que incluyen la pérdida completa de un sitio.

Los cambios pequeños en las políticas de protección y recuperación de datos pueden tener un efecto significativo en la arquitectura general de almacenamiento, respaldo y recuperación. Es crucial definir y documentar los estándares antes de comenzar a trabajar de diseño, para evitar complicar la arquitectura de protección de datos. Las funciones o niveles de protección innecesarios generan costes innecesarios y gastos generales de gestión, y un requisito que al principio se pasa por alto puede dirigir un proyecto en la dirección equivocada o requerir cambios de diseño de última hora.

Objetivo de tiempo de recuperación

El objetivo de tiempo de recuperación (RTO) define el tiempo máximo permitido para la recuperación de un servicio. Por ejemplo, una base de datos de recursos humanos podría tener un objetivo de tiempo de recuperación de 24 horas porque, si bien sería un inconveniente perder el acceso a estos datos durante la jornada laboral, la empresa aún puede seguir funcionando. Por el contrario, una base de datos que respalde el libro mayor general de un banco tendría un RTO medido en minutos o incluso segundos. Un RTO de cero no es posible, porque debe haber una manera de diferenciar entre una interrupción real del servicio y un evento rutinario, como un paquete de red perdido. Sin embargo, un objetivo de tiempo de recuperación de casi cero es un requisito típico.

Objetivo de punto de recuperación

El objetivo de punto de recuperación (RPO) define la pérdida de datos máxima tolerable. En muchos casos, el objetivo de punto de recuperación solo viene determinado por la frecuencia de las copias Snapshot o las actualizaciones de snapmirror.

En algunos casos, el objetivo de punto de recuperación puede hacerse más agresivo ya que protege de forma selectiva ciertos datos con mayor frecuencia. En un contexto de base de datos, el RPO suele ser una cuestión de cuántos datos de registro se pueden perder en una situación específica. En un escenario típico de recuperación en el que una base de datos está dañada debido a un error de producto o de usuario, el RPO debe ser cero, lo que significa que no debe haber pérdida de datos. El procedimiento de recuperación implica restaurar una copia anterior de los archivos de base de datos y, a continuación, volver a reproducir los archivos de registro para que el estado de la base de datos alcance el momento deseado. Los archivos de registro necesarios para esta operación ya deben estar en su lugar en la ubicación original.

En escenarios inusuales, los datos de registro pueden perderse. Por ejemplo, un ataque accidental o malintencionado `rm -rf *` de archivos de base de datos podría resultar en la eliminación de todos los datos. La única opción sería restaurar desde la copia de seguridad, incluidos los archivos de registro, y algunos datos inevitablemente se perderían. La única opción para mejorar el RPO en un entorno de backup tradicional sería realizar backups repetidos de los datos de registro. Sin embargo, esto tiene limitaciones debido al movimiento constante de datos y la dificultad de mantener un sistema de backup como un servicio en constante ejecución. Una de las ventajas de los sistemas de almacenamiento avanzados es la capacidad de proteger los datos frente a daños accidentales o malintencionados en los archivos para proporcionar, de este modo, un mejor objetivo de punto de recuperación sin transferir datos.

Recuperación tras siniestros

La recuperación tras desastres incluye la arquitectura de TI, las políticas y los procedimientos necesarios para recuperar un servicio en caso de desastre físico. Esto puede incluir inundaciones, incendios o personas que actúen con intención maliciosa o negligente.

La recuperación ante desastres va más allá de un conjunto de procedimientos de recuperación. Se trata del proceso completo de identificar los diversos riesgos, de definir los requisitos de recuperación de datos y

continuidad del servicio, y de proporcionar la arquitectura correcta con los procedimientos asociados.

Cuando se establecen requisitos de protección de datos, es fundamental diferenciar entre los requisitos típicos de RPO y RTO, así como los requisitos de RPO y RTO necesarios para la recuperación ante desastres. Algunos entornos de aplicaciones requieren un objetivo de punto de recuperación de cero y un objetivo de tiempo de recuperación de casi cero para situaciones de pérdida de datos, que van desde un error relativamente normal del usuario hasta un incendio que destruya un centro de datos. Sin embargo, estos altos niveles de protección tienen consecuencias administrativas y de costes.

En general, los requisitos de recuperación de datos sin desastre deben ser estrictos por dos motivos. En primer lugar, los errores en las aplicaciones y los errores de los usuarios que dañan los datos son previsibles hasta el punto de que son casi inevitables. En segundo lugar, no es difícil diseñar una estrategia de backup que proporcione un RPO de cero y un RTO bajo, siempre que el sistema de almacenamiento no esté destruido. No hay motivo para no abordar un riesgo significativo que sea fácil de solucionar, por lo que los objetivos de RPO y RTO para la recuperación local deben ser agresivos.

Los requisitos del objetivo de tiempo de recuperación ante desastres y del objetivo de punto de recuperación varían mucho más según la probabilidad de que se produzca un desastre y las consecuencias de la pérdida de datos o las interrupciones de un negocio. Los requisitos del objetivo de punto de recuperación y del objetivo de tiempo de recuperación deben basarse en las necesidades reales de la empresa, no en los principios generales. Deben explicar múltiples escenarios de desastre lógicos y físicos.

Desastres lógicos

Entre los desastres lógicos se encuentra la corrupción de datos causada por los usuarios, errores de la aplicación o del SO y mal funcionamiento del software. Los desastres lógicos también pueden incluir ataques maliciosos de terceros con virus o gusanos, o mediante la explotación de las vulnerabilidades de las aplicaciones. En estos casos, la infraestructura física permanece intacta, pero los datos subyacentes ya no son válidos.

Un tipo cada vez más común de desastre lógico se conoce como ransomware, en el que se utiliza un vector de ataque para cifrar los datos. El cifrado no daña los datos, pero no los hace disponibles hasta que se realiza el pago a un tercero. Un número cada vez mayor de empresas se dirigen específicamente a ataques de ransomware. Para esta amenaza, NetApp ofrece copias Snapshot a prueba de manipulaciones donde ni siquiera el administrador de almacenamiento puede cambiar los datos protegidos antes de la fecha de caducidad configurada.

Desastres físicos

Los desastres físicos incluyen la falla de los componentes de una infraestructura que supera sus capacidades de redundancia y dan lugar a una pérdida de datos o una prolongada pérdida de servicio. Por ejemplo, la protección RAID proporciona redundancia de unidades de disco y el uso de HBA proporciona redundancia de puertos FC y cables FC. Los errores de hardware de dichos componentes son previsibles y no afectan a la disponibilidad.

En un entorno empresarial, generalmente es posible proteger la infraestructura de todo un sitio con componentes redundantes hasta el punto en que el único escenario de desastre físico previsible es la pérdida completa del sitio. En ese caso, el plan de la recuperación ante desastres depende de la replicación entre sitios.

Protección de datos síncrona y asíncrona

En un mundo ideal, todos los datos se replicarían de forma sincrónica en sitios dispersos geográficamente. Dicha replicación no siempre es factible o incluso posible por varias razones:

- La replicación síncrona aumenta inevitablemente la latencia de escritura porque todos los cambios deben replicarse en ambas ubicaciones antes de que la aplicación o base de datos pueda continuar con el procesamiento. El efecto sobre el rendimiento resultante es a veces inaceptable, lo que descarta el uso del mirroring síncrono.
- Al aumentar la adopción del almacenamiento SSD del 100 %, es más probable que se note latencia de escritura adicional, ya que las expectativas de rendimiento incluyen cientos de miles de IOPS y latencia inferior al milisegundo. Para obtener todas las ventajas del uso del 100 % de las unidades SSD es necesario volver a analizar la estrategia de recuperación ante desastres.
- Los conjuntos de datos siguen creciendo en términos de bytes, generando retos que exigen un ancho de banda suficiente para sostener la replicación síncrona.
- Los conjuntos de datos también crecen en términos de complejidad, lo que genera retos con la gestión de la replicación síncrona a gran escala.
- Las estrategias basadas en cloud a menudo implican mayores distancias de replicación y latencia, lo que excluye aún más el uso del mirroring síncrono.

NetApp ofrece soluciones que incluyen replicación sincrónica para las exigencias de recuperación de datos más exigentes y soluciones asincrónicas que permiten un mejor rendimiento y flexibilidad. Además, la tecnología de NetApp se integra sin problemas con muchas soluciones de replicación de terceros, como Oracle DataGuard

Tiempo de retención

El aspecto final de una estrategia de protección de datos es el tiempo de retención, que puede variar drásticamente.

- Normalmente, se requieren 14 días de backups nocturnos en el sitio principal y 90 días de backups almacenados en un sitio secundario.
- Muchos clientes crean archivos trimestrales independientes almacenados en diferentes medios.
- Es posible que una base de datos constantemente actualizada no necesite datos históricos y que las copias de seguridad solo se conserven durante unos pocos días.
- Los requisitos normativos pueden requerir la capacidad de recuperación hasta el punto de cualquier transacción arbitraria en un periodo de 365 días.

Disponibilidad de bases de datos

ONTAP se ha diseñado para ofrecer la máxima disponibilidad de las bases de datos de Oracle. Este documento no incluye una descripción completa de las funciones de alta disponibilidad de ONTAP. Sin embargo, al igual que sucede con la protección de datos, un conocimiento básico de esta funcionalidad es importante cuando se diseña una infraestructura de base de datos.

Parejas de HA

La unidad básica de alta disponibilidad es el par de alta disponibilidad. Cada pareja contiene enlaces redundantes para admitir la replicación de datos hacia NVRAM. NVRAM no es una caché de escritura. La RAM dentro de la controladora funciona como caché de escritura. El objetivo de la NVRAM es registrar temporalmente los datos como protección frente a un fallo inesperado del sistema. En este sentido, es similar a un redo log de base de datos.

Tanto la NVRAM como un redo log de base de datos se utilizan para almacenar datos rápidamente, lo que

permite que los cambios en los datos se confirmen lo más rápidamente posible. La actualización de los datos persistentes en las unidades (o archivos de datos) no se realiza hasta más adelante durante un proceso denominado punto de control en las plataformas ONTAP y en la mayoría de las bases de datos. Ni los datos de NVRAM ni los registros de recuperación de bases de datos se leen durante las operaciones normales.

Si una controladora falla abruptamente, es posible que existan cambios pendientes almacenados en la NVRAM que aún no se hayan escrito en las unidades. La controladora asociada detecta el fallo, toma el control de las unidades y aplica los cambios requeridos que se han almacenado en NVRAM.

Toma de control y retorno al nodo primario

La toma de control y la devolución hace referencia al proceso de transferencia de la responsabilidad de los recursos de almacenamiento entre los nodos de un par de alta disponibilidad. La toma de control y el retorno al nodo primario tienen dos aspectos:

- Gestión de la conectividad de red que permite el acceso a las unidades
- Gestión de las unidades en sí

Las interfaces de red que admiten el tráfico CIFS y NFS están configuradas tanto con un directorio raíz como con una ubicación de recuperación tras fallos. Una toma de control incluye mover las interfaces de red a su directorio raíz temporal en una interfaz física ubicada en las mismas subredes que la ubicación original. Un retorno primario incluye mover las interfaces de red de vuelta a sus ubicaciones originales. El comportamiento exacto se puede ajustar según sea necesario.

Las interfaces de red que admiten protocolos de bloques SAN como iSCSI y FC no se reubican durante la toma de control y el retorno al nodo primario. En su lugar, los LUN se deben aprovisionar con rutas que incluyan un par de HA completo, lo que da como resultado una ruta primaria y una secundaria.

> También se pueden configurar rutas adicionales a controladoras adicionales para admitir la reubicación de datos entre nodos de un clúster más grande, pero esto no forma parte del proceso de alta disponibilidad.

El segundo aspecto de la toma de control y la restauración es la transferencia de la propiedad del disco. El proceso exacto depende de múltiples factores, incluyendo la razón de la toma de control/devolución y las opciones de la línea de comandos emitidas. El objetivo es realizar la operación de la manera más eficiente posible. Aunque parezca que el proceso general requiera varios minutos, el momento en el que la propiedad de la unidad se realiza la transición de nodo a nodo generalmente se puede medir en segundos.

Tiempo de toma de control

El host de I/O experimenta una breve pausa en I/O durante operaciones de toma de control y devolución; pero no debe producirse una interrupción en las aplicaciones en un entorno configurado correctamente. El proceso de transición real en el que se demora I/O suele medirse en segundos, pero el host puede requerir más tiempo para reconocer el cambio en las rutas de datos y volver a enviar las operaciones de I/O.

La naturaleza de la interrupción depende del protocolo:

- Una interfaz de red que admite problemas de tráfico NFS y CIFS una solicitud de Protocolo de resolución de direcciones (ARP) a la red después de la transición hacia una nueva ubicación física. Esto hace que los conmutadores de red actualicen sus tablas de direcciones de control de acceso a medios (MAC) y reanuden el procesamiento de E/S. Las interrupciones en el caso de toma de control y devolución planificadas suelen medirse en segundos y, en muchos casos, no se pueden detectar. Puede que algunas redes sean más lentas para reconocer completamente el cambio en la ruta de red y algunos sistemas operativos pueden poner en cola muchas E/S en muy poco tiempo que deben reintentarse. Esto puede

ampliar el tiempo necesario para reanudar la actividad de I/O.

- Una interfaz de red que admite protocolos SAN no realiza la transición a una nueva ubicación. Un SO host debe cambiar la ruta o las rutas en uso. La pausa en I/O observada por el host depende de varios factores. Desde el punto de vista de un sistema de almacenamiento, el período en el que no se puede ofrecer I/O es solo unos segundos. Sin embargo, los sistemas operativos de host diferentes pueden requerir más tiempo para permitir que se agote el tiempo de espera de una E/S antes de volver a intentarlo. Los sistemas operativos más nuevos son más capaces de reconocer un cambio de ruta mucho más rápido, pero los sistemas operativos más antiguos normalmente requieren hasta 30 segundos para reconocer un cambio.

En la siguiente tabla, se muestran los tiempos de toma de control esperados durante el que el sistema de almacenamiento no puede ofrecer datos a un entorno de aplicación. No debe haber ningún error en ningún entorno de aplicación, la toma de control debería aparecer como una breve pausa en el procesamiento de E/S.

	NFS	AFF	ASA
Toma de control planificada	15 seg	6-10 seg	2-3 seg
Respaldo no planificado	30 seg	6-10 seg	2-3 seg

Sumas de comprobación e integridad de los datos

ONTAP y sus protocolos admitidos incluyen varias funciones que protegen la integridad de las bases de datos de Oracle, incluidos los datos en reposo y la transmisión de datos a través de la red.

La protección de datos lógicos en ONTAP consta de tres requisitos clave:

- Los datos deben protegerse contra la corrupción de datos.
- Los datos deben protegerse contra un fallo de unidad.
- Los cambios en los datos deben protegerse contra la pérdida.

Estas tres necesidades se tratan en las siguientes secciones.

Corrupción de la red: Sumas de comprobación

El nivel más básico de protección de datos es la suma de comprobación, que es un código especial de detección de errores almacenado junto con los datos. La corrupción de datos durante la transmisión de red se detecta con el uso de una suma de comprobación y, en algunos casos, varias sumas de comprobación.

Por ejemplo, una trama de FC incluye una forma de suma de comprobación denominada comprobación de redundancia cíclica (CRC) para asegurarse de que la carga útil no está dañada en tránsito. El transmisor envía tanto los datos como el CRC de los datos. El receptor de una trama FC vuelve a calcular el CRC de los datos recibidos para asegurarse de que coincida con el CRC transmitido. Si el CRC recién calculado no coincide con el CRC conectado a la trama, los datos están dañados y se descarta o rechaza la trama de FC. Las operaciones de I/O iSCSI incluyen sumas de comprobación en las capas TCP/IP y Ethernet y, para una protección adicional, también se puede incluir protección CRC opcional en la capa SCSI. Cualquier daño de bit en el cable se detecta mediante la capa TCP o la capa IP, lo que provoca la retransmisión del paquete. Al igual que con FC, los errores en el CRC de SCSI provocan un descarte o el rechazo de la operación.

Daños en unidades: Sumas de comprobación

También se utilizan sumas de comprobación para verificar la integridad de los datos almacenados en las unidades. Los bloques de datos escritos en las unidades se almacenan con una función de suma de comprobación que genera un número impredecible ligado a los datos originales. Cuando se leen datos de la unidad, la suma de comprobación se vuelve a calcular y se compara con la suma de comprobación almacenada. Si no coincide, los datos se han dañado y deben ser recuperados por la capa RAID.

Datos dañados: Escrituras perdidas

Uno de los tipos de daños más difíciles de detectar es una escritura perdida o ubicada incorrectamente. Cuando se reconoce una escritura, se debe escribir en el soporte en la ubicación correcta. Los datos dañados in situ son relativamente fáciles de detectar usando una sencilla suma de comprobación almacenada con los datos. Sin embargo, si la escritura simplemente se pierde, es posible que aún exista la versión anterior de los datos y la suma de comprobación sea correcta. Si la escritura se realiza en una ubicación física incorrecta, la suma de comprobación asociada sería una vez más válida para los datos almacenados, aunque la escritura haya destruido otros datos.

La solución a este reto es la siguiente:

- Una operación de escritura debe incluir metadatos que indiquen la ubicación donde se espera que se encuentre la escritura.
- Una operación de escritura debe incluir algún tipo de identificador de versión.

Cuando ONTAP escribe un bloque, incluye los datos donde pertenece el bloque. Si una lectura posterior identifica un bloque, pero los metadatos indican que pertenece a la ubicación 123 cuando se encontró en la ubicación 456, la escritura se ha colocado de forma incorrecta.

Detectar una escritura totalmente perdida es más difícil. La explicación es muy complicada, pero básicamente ONTAP almacena los metadatos de manera que una operación de escritura da como resultado actualizaciones en dos ubicaciones distintas en las unidades. Si se pierde una escritura, una lectura posterior de los datos y los metadatos asociados muestra dos identidades de versión diferentes. Esto indica que la unidad no completó la escritura.

Los daños en la escritura perdidos o mal ubicados son extremadamente raros, pero, a medida que las unidades siguen creciendo y los conjuntos de datos pasan a la escala de exabytes, el riesgo aumenta. La detección de escritura perdida debe incluirse en cualquier sistema de almacenamiento que admita cargas de trabajo de base de datos.

Fallos de unidad: RAID, RAID DP y RAID-TEC

Si se detecta que un bloque de datos en una unidad está dañado, o que toda la unidad falla y no está totalmente disponible, los datos deben reconstituirse. Esto se realiza en ONTAP utilizando unidades de paridad. Los datos se dividen entre varias unidades de datos y, a continuación, se generan datos de paridad. Se almacena por separado de los datos originales.

ONTAP utilizó originalmente RAID 4, que utiliza una sola unidad de paridad para cada grupo de unidades de datos. El resultado fue que cualquier unidad del grupo podría fallar sin producir una pérdida de datos. Si se produjo un error en la unidad de paridad, no se dañaron los datos y se pudo construir una nueva unidad de paridad. Si falla una unidad de datos única, las unidades restantes podrían usarse con la unidad de paridad para volver a generar los datos ausentes.

Cuando las unidades eran pequeñas, la posibilidad estadística de que fallaran en dos unidades a la vez era insignificante. A medida que aumenta la capacidad de las unidades, también aumenta el tiempo necesario para reconstruir los datos tras un fallo de unidad. Esto ha aumentado el intervalo en el que un segundo fallo

de unidad provocaría la pérdida de datos. Además, el proceso de recompilación crea una gran cantidad de I/O adicionales en las unidades supervivientes. A medida que las unidades envejecen, también aumenta el riesgo de la carga adicional que produce un segundo fallo de unidad. Por último, incluso si el riesgo de pérdida de datos no aumentara con el uso continuado de RAID 4, las consecuencias de la pérdida de datos serían más graves. Cuantos más datos se pierdan en caso de un fallo de un grupo RAID, más tiempo se necesitaría para recuperar los datos, lo que prolonga la interrupción del negocio.

Estos problemas llevaron a NetApp a desarrollar la tecnología NetApp RAID DP, una variante de RAID 6. Esta solución incluye dos unidades de paridad, lo que significa que dos unidades cualesquiera de un grupo RAID pueden fallar sin crear pérdida de datos. El tamaño de las unidades ha continuado creciendo, lo que finalmente llevó a NetApp a desarrollar la tecnología NetApp RAID-TEC, que introduce una tercera unidad de paridad.

Algunas mejores prácticas históricas de bases de datos recomiendan el uso de RAID-10, también conocido como mirroring segmentado. Esto ofrece menos protección de datos que RAID DP, ya que existen varias situaciones de fallo de dos discos, mientras que en RAID DP no hay ninguna.

También hay algunas mejores prácticas históricas de bases de datos que indican que se prefiere RAID-10 a las opciones de RAID-4/5/6 debido a cuestiones de rendimiento. En ocasiones, estas recomendaciones se refieren a una penalización de RAID. Aunque estas recomendaciones son generalmente correctas, no son aplicables a las implementaciones de RAID en ONTAP. El problema de rendimiento está relacionado con la regeneración de paridad. Con las implementaciones de RAID tradicionales, procesar las escrituras aleatorias rutinarias realizadas por una base de datos requiere varias lecturas de disco para regenerar los datos de paridad y completar la escritura. La penalización se define como las IOPS de lectura adicional necesarias para ejecutar operaciones de escritura.

ONTAP no incurre en una penalización de RAID, ya que las escrituras se almacenan en memoria donde se genera la paridad y se escriben en el disco como una única franja de RAID. No se requieren lecturas para completar la operación de escritura.

En resumen, en comparación con RAID 10, RAID DP y RAID-TEC ofrecen mucha más capacidad utilizable, una mejor protección ante fallos de unidad y sin sacrificios de rendimiento.

Protección contra fallos del hardware: NVRAM

Cualquier cabina de almacenamiento que sirva a una carga de trabajo de base de datos debe procesar operaciones de escritura lo más rápido posible. Además, una operación de escritura debe protegerse contra pérdidas provocadas por eventos inesperados, como un fallo de alimentación. Esto significa que cualquier operación de escritura debe almacenarse de forma segura en al menos dos ubicaciones.

Los sistemas AFF y FAS confían en NVRAM para cumplir estos requisitos. El proceso de escritura funciona de la siguiente manera:

1. Los datos de escritura entrantes se almacenan en la RAM.
2. Los cambios que se deben realizar en los datos del disco se registran en NVRAM en el nodo local y el asociado. NVRAM no es una caché de escritura, sino un diario similar a un redo log de base de datos. En condiciones normales, no se lee. Solo se utiliza para recuperación, como después de un fallo de alimentación durante el procesamiento de I/O.
3. A continuación, la escritura se reconoce en el host.

El proceso de escritura en esta fase se completa desde el punto de vista de la aplicación y los datos están protegidos contra pérdidas debido a que están almacenados en dos ubicaciones diferentes. Eventualmente, los cambios se escriben en el disco, pero este proceso es fuera de banda desde el punto de vista de la aplicación, porque se produce una vez que se reconoce la escritura y, por lo tanto, no afecta a la latencia. Este

proceso es una vez más similar al registro de la base de datos. Un cambio en la base de datos se registra en los redo logs lo antes posible y el cambio se confirma como confirmado. Las actualizaciones de los archivos de datos se producen mucho más tarde y no afectan directamente a la velocidad de procesamiento.

En caso de que se produzca un fallo en la controladora, la controladora asociada toma la propiedad de los discos necesarios y reproduce los datos registrados en la NVRAM para recuperar las operaciones de I/O que estuvieran en curso al producirse el fallo.

Protección contra fallos de hardware: 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 con el partner de alta disponibilidad falla, estos datos en curso ya no se duplicarán.

Si la NVRAM local informa de un error, el nodo se apaga. Este apagado hace que se produzca una conmutación al nodo de respaldo con una controladora asociada de alta disponibilidad. No se pierden datos porque la controladora que experimenta el fallo no reconoció la operación de escritura.

ONTAP no permite una conmutación por error cuando los datos no están sincronizados a menos que se vean obligados a recurrir a la conmutación por error. 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 porque las bases de datos mantienen grandes cachés internos de datos en el disco. Si se produce una conmutación por error forzada, los cambios previamente aceptados 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 datos contra esta situación, ONTAP permite configurar volúmenes para una 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 una aplicación para que no utilicen datos obsoletos. No se deben perder los datos porque debe haber alguna escritura reconocida en la cabina de almacenamiento.

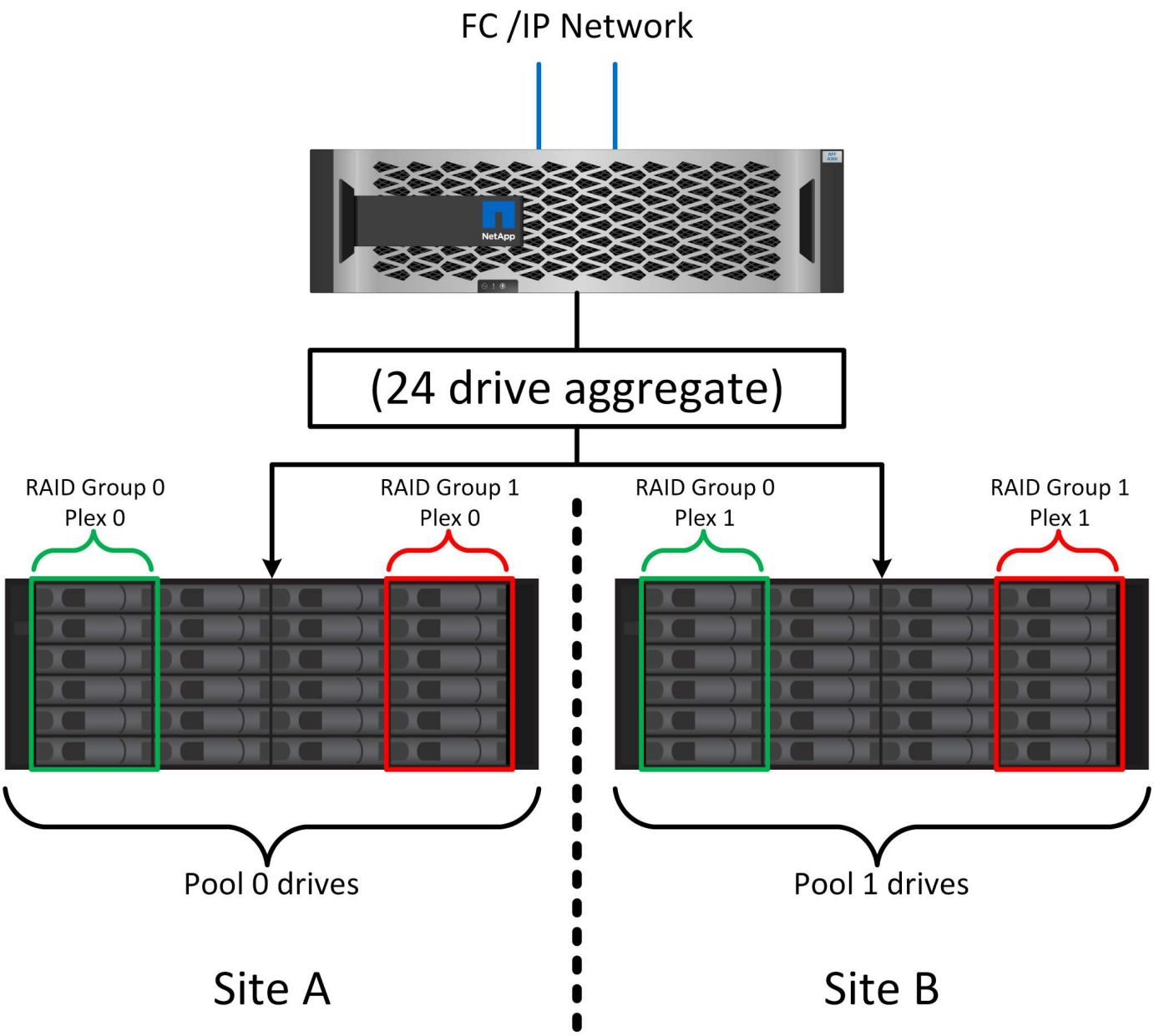
Los siguientes pasos habituales son para que un administrador apague completamente los hosts antes de volver a poner manualmente los LUN y los volúmenes de nuevo en línea. Aunque estos pasos pueden implicar cierto trabajo, este enfoque es la manera más segura de garantizar la integridad de los datos.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

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:

- 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.
- 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.
- 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 del mirroring. Esto proporciona una protección de datos continua, 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. RAID Group 0 incluye un plex de 6 unidades en el sitio A duplicado en un plex de 6 unidades en el sitio B. Del mismo modo, RAID Group 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 alimentados por fuentes de alimentación independientes de sistemas UPS independientes.

Sumas de comprobación

El tema de las sumas de comprobación es de particular interés para los administradores de bases de datos que están acostumbrados a usar backups en streaming de Oracle RMAN, que migran a backups basados en instantáneas. Una función de RMAN es que realiza comprobaciones de integridad durante las operaciones de copia de seguridad. Aunque esta función posee cierto valor, su principal ventaja es en una base de datos que no se utiliza en una cabina de almacenamiento moderna. Cuando se utilizan unidades físicas en una base de datos de Oracle, resulta casi seguro que los daños eventualmente se producen cuando las unidades envejecen, un problema que resuelven las sumas de comprobación basadas en cabinas de almacenamiento reales.

Con una cabina de almacenamiento real, la integridad de los datos se protege utilizando sumas de comprobación en varios niveles. Si los datos están dañados en una red basada en IP, la capa Protocolo de control de transmisión (TCP) rechaza los datos del paquete y solicita la retransmisión. El protocolo FC incluye sumas de comprobación, al igual que los datos SCSI encapsulados. Después de que se encuentra en la cabina, ONTAP tiene protección RAID y suma de comprobación. La corrupción puede ocurrir, pero, como en la mayoría de las matrices empresariales, se detecta y corrige. Normalmente, falla una unidad completa, solicita una reconstrucción de RAID y la integridad de la base de datos no se ve afectada. Todavía es posible que los bytes individuales en una unidad sean dañados por la radiación cósmica o las células flash que fallan. Si esto sucede, se producirá un error en la comprobación de paridad, se producirá un error en la unidad y se iniciará la recompilación de RAID. Una vez más, la integridad de los datos no se ve afectada. La última línea de defensa es el uso de sumas de control. Si, por ejemplo, un error catastrófico de firmware en una unidad daña datos que de algún modo no se detectó mediante una comprobación de paridad de RAID, la suma de comprobación no coincidiría y ONTAP evitaría la transferencia de un bloque dañado antes de que la base de datos de Oracle pudiera recibirlo.

La arquitectura de archivo de datos y redo log de Oracle también está diseñada para ofrecer el nivel más alto posible de integridad de datos, incluso en circunstancias extremas. En el nivel más básico, los bloques de Oracle incluyen suma de comprobación y comprobaciones lógicas básicas con casi todas las E/S. Si Oracle no se ha bloqueado o ha puesto un tablespace fuera de línea, los datos estarán intactos. El grado de comprobación de la integridad de los datos es ajustable y Oracle también puede configurarse para confirmar las escrituras. Como resultado, casi todos los escenarios de accidente y fallo se pueden recuperar, y en el caso extremadamente raro de una situación irrecuperable, la corrupción se detecta rápidamente.

La mayoría de los clientes de NetApp que utilizan bases de datos Oracle interrumpen el uso de RMAN y otros productos de backup después de la migración a backups basados en snapshots. Todavía hay opciones en las que se puede utilizar RMAN para realizar la recuperación a nivel de bloque con SnapCenter. Sin embargo, en el día a día, RMAN, NetBackup y otros productos sólo se utilizan ocasionalmente para crear copias de archivado mensuales o trimestrales.

Algunos clientes eligen correr `dbv` periódicamente para realizar comprobaciones de integridad de sus bases de datos existentes. NetApp desaconseja esta práctica porque crea una carga de I/O innecesaria. Como se mencionó anteriormente, si la base de datos no estaba experimentando problemas anteriormente, la posibilidad de `dbv` La detección de un problema es cercana a cero, y esta utilidad crea una carga secuencial de I/O muy elevada en la red y el sistema de almacenamiento. A menos que exista un motivo para creer que existe corrupción, como la exposición a un bug de Oracle conocido, no hay motivo para ejecutarse `dbv`.

Conceptos básicos de backup y recuperación

Backups basados en Snapshot

La base de la protección de datos de bases de datos de Oracle en ONTAP es la tecnología Snapshot de NetApp.

Los valores clave son los siguientes:

- **Simplicidad.** Una instantánea es una copia de solo lectura del contenido de un contenedor de datos en un momento específico.
- **Eficiencia.** Las instantáneas no requieren espacio en el momento de la creación. El espacio solo se consume cuando se modifican los datos.
- **Capacidad de gestión.** Una estrategia de copia de seguridad basada en instantáneas es fácil de configurar y administrar porque las instantáneas son una parte nativa del sistema operativo de almacenamiento. Si el sistema de almacenamiento está encendido, está listo para crear backups.
- **Escalabilidad.** Se pueden conservar hasta 1024 copias de seguridad de un único contenedor de archivos y LUN. En el caso de conjuntos de datos complejos, es posible proteger varios contenedores de datos con un único conjunto coherente de copias Snapshot.
- El rendimiento no se ve afectado, independientemente de que un volumen contenga 1024 snapshots o ninguna.

Aunque muchos proveedores de almacenamiento ofrecen tecnología Snapshot, la tecnología Snapshot dentro de ONTAP es única y ofrece beneficios importantes para los entornos de aplicaciones y bases de datos empresariales:

- Las copias Snapshot forman parte del sistema de archivos WAFL (Write-Anywhere File Layout) subyacente. No son una tecnología complementaria ni externa. Esto simplifica la gestión, ya que el sistema de almacenamiento es el sistema de backup.
- Las copias Snapshot no afectan al rendimiento, a excepción de algunos casos periféricos como cuando se almacenan tantos datos en copias snapshot que el sistema de almacenamiento subyacente llena.
- El término «grupo de coherencia» se utiliza a menudo para referirse a una agrupación de objetos de almacenamiento que se gestionan como una colección consistente de datos. Una Snapshot de un volumen ONTAP determinado constituye un backup de grupo de coherencia.

Las copias Snapshot de ONTAP también ofrecen una escalabilidad mejor que la tecnología de la competencia. Los clientes pueden almacenar 5, 50 o 500 copias Snapshot sin que esto afecte al rendimiento. El número máximo de snapshots que se permite actualmente en un volumen es 1024. Si se requiere más retención de instantáneas, existen opciones para configurar las instantáneas en cascada a volúmenes adicionales.

Como resultado, proteger un conjunto de datos alojado en ONTAP es sencillo y altamente escalable. Los backups no requieren el traslado de datos, por lo que puede adaptarse a las necesidades del negocio en lugar de a las limitaciones de las tasas de transferencia de red, un gran número de unidades de cinta o áreas de almacenamiento provisional de discos.

¿Una snapshot es un backup?

Una pregunta frecuente acerca del uso de las copias Snapshot como estrategia de protección de datos es el hecho de que los datos «reales» y los datos de copias Snapshot se encuentran en las mismas unidades. La pérdida de esas unidades provocaría la pérdida de los datos primarios y el backup.

Este es un problema válido. Los snapshots locales se usan para necesidades de backup y recuperación diarias y, en ese sentido, la snapshot es un backup. Cerca del 99 % de todos los escenarios de recuperación en entornos NetApp utilizan copias Snapshot para satisfacer incluso los requisitos de objetivo de tiempo de recuperación más agresivos.

Sin embargo, las copias Snapshot locales nunca deberían ser la única estrategia de backup, por lo que NetApp ofrece tecnología como la replicación de SnapMirror y SnapVault para replicar de forma rápida y eficiente copias Snapshot en un conjunto de unidades independiente. En una solución correctamente

diseñada con copias Snapshot y replicación Snapshot, el uso de la cinta puede minimizarse tal vez a un archivo trimestral o eliminarse totalmente.

Backups basados en Snapshot

Existen muchas opciones para usar las copias Snapshot de ONTAP para proteger los datos, y las copias Snapshot son la base de muchas otras funciones de ONTAP, como replicación, recuperación ante desastres y clonación. Una descripción completa de la tecnología de instantáneas está fuera del alcance de este documento, pero en las siguientes secciones se proporciona una descripción general.

Existen dos métodos principales para crear una copia Snapshot de un conjunto de datos:

- Backups coherentes con los fallos
- Backups para aplicaciones

Un backup coherente con los fallos de un conjunto de datos hace referencia a la captura de toda la estructura del conjunto de datos en un único punto de tiempo. Si el conjunto de datos se almacena en un único volumen, el proceso es sencillo; se puede crear una copia Snapshot en cualquier momento. Si un conjunto de datos abarca volúmenes, es necesario crear una snapshot de grupo de coherencia (CG). Existen varias opciones para crear snapshots de CG, como el software NetApp SnapCenter, funciones nativas del grupo de coherencia ONTAP y scripts que se mantienen por el usuario.

Los backups coherentes con los fallos se utilizan principalmente cuando la recuperación punto del backup es suficiente. Cuando se necesita una recuperación más granular, por lo general se necesitan backups coherentes con las aplicaciones.

A menudo, la palabra «consistente» en «coherente con las aplicaciones» resulta una denominación errónea. Por ejemplo, colocar una base de datos de Oracle en modo de backup se denomina backup coherente con las aplicaciones, pero los datos no se hacen coherentes ni se ponen en modo inactivo de ninguna forma. Los datos siguen cambiando durante el backup. Por el contrario, la mayoría de los backups de MySQL y Microsoft SQL Server realmente ralentizan los datos antes de ejecutar el backup. VMware puede o no hacer que ciertos archivos sean consistentes.

Grupos de consistencia

El término «grupo de coherencia» hace referencia a la capacidad de una cabina de almacenamiento para gestionar varios recursos de almacenamiento como una sola imagen. Por ejemplo, una base de datos puede consistir en 10 LUN. La cabina debe ser capaz de realizar backup, restaurar y replicar esos 10 LUN de forma coherente. La restauración no es posible si las imágenes de las LUN no eran consistentes en el punto de backup. Para replicar estos 10 LUN es necesario que todas las réplicas estén perfectamente sincronizadas entre sí.

El término «grupo de coherencia» no se utiliza con frecuencia cuando se habla de ONTAP, porque la coherencia siempre ha sido una función básica del volumen y de la arquitectura de agregado en ONTAP. Muchas otras cabinas de almacenamiento gestionan LUN o sistemas de archivos como unidades individuales. Podrían configurarse opcionalmente como «grupo de consistencia» para fines de protección de datos, pero este es un paso adicional en la configuración.

ONTAP siempre ha podido capturar imágenes de datos replicadas y locales coherentes. Aunque los distintos volúmenes de un sistema ONTAP no suelen describirse formalmente como un grupo de coherencia, eso es lo que son. Una copia Snapshot de ese volumen es una imagen de grupo de coherencia, la restauración de esa copia Snapshot es una restauración de grupo de coherencia, y tanto SnapMirror como SnapVault ofrecen replicación de grupo de coherencia.

Snapshots de grupo de coherencia

Las snapshots de grupo de consistencia (cg-snapshots) son una extensión de la tecnología Snapshot básica de ONTAP. Una operación Snapshot estándar crea una imagen coherente de todos los datos dentro de un único volumen, pero a veces es necesario crear un conjunto coherente de instantáneas en varios volúmenes e incluso entre varios sistemas de almacenamiento. El resultado es un conjunto de instantáneas que se pueden utilizar de la misma manera que una instantánea de un solo volumen individual. Se pueden utilizar para la recuperación de datos locales, replicar para la recuperación ante desastres o clonar como una única unidad coherente.

El mayor uso conocido de cg-snapshots es para un entorno de base de datos de aproximadamente 1PB GB de tamaño que abarca 12 controladoras. Las cg-snapshots creadas en este sistema se han utilizado para backup, recuperación y clonado.

La mayoría de las veces, cuando un conjunto de datos abarca volúmenes y se debe conservar el orden de escritura, el software de gestión elegido utiliza automáticamente una instantánea de cg. No es necesario comprender los detalles técnicos de cg-snapshots en estos casos. No obstante, hay situaciones en las que los complejos requisitos de protección de datos requieran un control detallado del proceso de protección y replicación de datos. Los flujos de trabajo de automatización o el uso de scripts personalizados para llamar a las API de cg-snapshot son algunas de las opciones. Para comprender la mejor opción y el rol de cg-snapshot se requiere una explicación más detallada de la tecnología.

La creación de un conjunto de cg-snapshots es un proceso de dos pasos:

1. Establezca el aislamiento de escritura en todos los volúmenes de destino.
2. Crear snapshots de dichos volúmenes mientras se encuentra en estado protegido.

El cercado de escritura se establece en serie. Esto significa que, a medida que se configura el proceso de barrera en varios volúmenes, las operaciones de I/O de escritura se congelan en el primer volumen de la secuencia, a medida que sigue confirmándose con los volúmenes que aparecen más adelante. Esto puede parecer que, en un principio, no cumple el requisito de conservación de la orden de escritura, pero eso solo se aplica a I/O que se emite de forma asíncrona en el host y no depende de ninguna otra escritura.

Por ejemplo, una base de datos puede emitir muchas actualizaciones de archivos de datos asíncronos y permitir que el sistema operativo vuelva a ordenar la I/O y completarlas de acuerdo con su propia configuración del programador. El orden de este tipo de I/O no se puede garantizar porque la aplicación y el sistema operativo ya han liberado el requisito de conservar el orden de escritura.

Como ejemplo de contador, la mayor parte de la actividad de registro de la base de datos es síncrona. La base de datos no continúa con más escrituras de registro hasta que se reconozca la E/S y se mantenga el orden de esas escrituras. Si un registro de I/O llega a un volumen cercado, no se reconoce y la aplicación se bloquea en otras escrituras. Del mismo modo, la I/O de metadatos del sistema de archivos suele ser síncrona. Por ejemplo, no se debe perder una operación de eliminación de archivos. Si un sistema operativo con un sistema de archivos xfs suprimió un archivo y la E/S que actualizó los metadatos del sistema de archivos xfs para eliminar la referencia a ese archivo aterrizó en un volumen cercado, la actividad del sistema de archivos se detendría. De este modo se garantiza la integridad del sistema de archivos durante las operaciones cg-snapshot.

Después de configurar el control de escritura en los volúmenes de destino, están listos para la creación de las copias Snapshot. No es necesario crear las copias Snapshot precisamente al mismo tiempo, ya que el estado de los volúmenes se congela desde un punto de vista de escritura dependiente. Para protegerse frente a un defecto en la aplicación que crea las copias cg-snapshots, la barrera de escritura inicial incluye un tiempo de espera configurable en el que ONTAP libera automáticamente la barrera y reanuda el procesamiento de escritura transcurridos un número de segundos definido. Si todas las Snapshot se crean antes de que se agote el tiempo de espera, el conjunto de snapshots resultante es un grupo de coherencia válido.

Orden de escritura dependiente

Desde un punto de vista técnico, la clave para un grupo de consistencia es preservar el orden de escritura y, específicamente, el orden de escritura dependiente. Por ejemplo, una base de datos que escribe en 10 LUN escribe simultáneamente en todas ellas. Muchas escrituras se emiten de forma asíncrona, por lo que el orden en que se completan no es importante y el orden en que se realizan varía según el comportamiento del sistema operativo y de la red.

Algunas operaciones de escritura deben estar presentes en el disco antes de que la base de datos pueda continuar con escrituras adicionales. Estas operaciones de escritura cruciales se denominan escrituras dependientes. La E/S de escritura posterior depende de la presencia de estas escrituras en el disco. Cualquier snapshot, recuperación o replicación de estas 10 LUN debe asegurarse de que la orden de escritura dependiente está garantizada. Las actualizaciones del sistema de archivos son otro ejemplo de escrituras dependientes del orden de escritura. El orden en el que se realizan los cambios en el sistema de archivos debe conservarse o todo el sistema de archivos podría dañarse.

Estrategias

Existen dos enfoques principales para los backups basados en Snapshot:

- Backups coherentes con los fallos
- Backups activos protegidos de Snapshot

Una copia de seguridad coherente con los fallos de una base de datos se refiere a la captura de toda la estructura de la base de datos, incluidos archivos de datos, redo logs y archivos de control, en un único punto en el tiempo. Si la base de datos se almacena en un único volumen, el proceso es sencillo; se puede crear una copia Snapshot en cualquier momento. Si una base de datos abarca volúmenes, debe crearse una snapshot de grupo de coherencia (CG). Existen varias opciones para crear snapshots de CG, como el software NetApp SnapCenter, funciones nativas del grupo de coherencia ONTAP y scripts que se mantienen por el usuario.

Los backups de Snapshot coherentes con los fallos se usan principalmente cuando es suficiente con la recuperación punto del backup. Los registros de archivos se pueden aplicar bajo ciertas circunstancias, pero cuando se requiere una recuperación puntual más granular, es preferible un backup online.

El procedimiento básico para un backup en línea basado en Snapshot es el siguiente:

1. Coloque la base de datos en `backup` modo.
2. Cree una instantánea de todos los volúmenes que alojan archivos de datos.
3. Salga `backup` modo.
4. Ejecute el comando `alter system archive log current` para forzar el archivado de registros.
5. Crear instantáneas de todos los volúmenes que alojan los archive logs.

Este procedimiento produce un juego de instantáneas que contienen archivos de datos en modo de backup y los archive logs críticos generados durante el modo de backup. Estos son los dos requisitos para recuperar una base de datos. Los archivos, como los archivos de control, también deben protegerse por conveniencia, pero el único requisito absoluto es la protección de los archivos de datos y los registros de archivos.

Aunque los diferentes clientes pueden tener estrategias muy diferentes, casi todas estas estrategias se basan en última instancia en los mismos principios descritos a continuación.

Recuperación basada en Snapshot

Al diseñar diseños de volúmenes para bases de datos Oracle, la primera decisión es si utilizar tecnología NetApp SnapRestore basada en volúmenes (VBSR).

El SnapRestore basado en volúmenes permite revertir un volumen casi instantáneamente a un momento específico anterior. Debido a que se revierten todos los datos del volumen, es posible que VBSR no sea apropiado para todos los casos de uso. Por ejemplo, si se almacena una base de datos completa, incluidos archivos de datos, registros de recuperación y registros de archivos, en un solo volumen y este volumen se restaura con VBSR, los datos se pierden porque se descartan los datos de archive log y redo más recientes.

VBSR no se requiere para la restauración. Muchas bases de datos pueden restaurarse utilizando SnapRestore de archivo único (SFSR) basado en archivos o simplemente copiando archivos del snapshot al sistema de archivos activo.

Se prefiere VBSR cuando una base de datos es muy grande o cuando se debe recuperar lo antes posible, y el uso de VBSR requiere aislamiento de los archivos de datos. En un entorno NFS, los archivos de datos de una base de datos determinada deben estar almacenados en volúmenes dedicados que no estén contaminados por ningún otro tipo de archivo. En un entorno SAN, los archivos de datos deben almacenarse en LUN dedicadas en volúmenes dedicados. Si se utiliza un gestor de volúmenes (incluido Oracle Automatic Storage Management [ASM]), el grupo de discos también debe estar dedicado a los archivos de datos.

El aislamiento de archivos de datos de esta manera permite que se reviertan a un estado anterior sin dañar otros sistemas de archivos.

Reserva de Snapshot

Para cada volumen con datos de Oracle en un entorno SAN, el `percent-snapshot-space` Debe establecerse en cero porque reservar espacio para una snapshot en un entorno de LUN no es útil. Si la reserva fraccionaria se establece en 100, una copia snapshot de un volumen con unidades lógicas requiere suficiente espacio libre en el volumen, excluida la reserva de snapshot, para absorber un 100% de renovación de todos los datos. Si la reserva fraccionaria se define en un valor menor, se requiere una cantidad de espacio libre correspondiente menor, pero siempre excluye la reserva de instantáneas. Esto significa que se desperdicia el espacio de reserva de snapshot en un entorno de LUN.

En un entorno NFS, hay dos opciones:

- Ajuste la `percent-snapshot-space` basado en el consumo de espacio esperado de la instantánea.
- Ajuste la `percent-snapshot-space` a cero y gestione el consumo de espacio activo y snapshot de forma colectiva.

Con la primera opción, `percent-snapshot-space` se establece en un valor distinto de cero, normalmente alrededor del 20%. Este espacio se oculta al usuario. Sin embargo, este valor no crea un límite de utilización. Si una base de datos con una reserva del 20% experimenta una rotación del 30%, el espacio de la instantánea puede crecer más allá de los límites de la reserva del 20% y ocupar espacio sin reservar.

La principal ventaja de establecer una reserva en un valor como 20% es verificar que algo de espacio esté siempre disponible para las instantáneas. Por ejemplo, un volumen de 1TB GB con una reserva del 20% solo permitiría que un administrador de bases de datos (DBA) almacene 800GB TB de datos. Esta configuración garantiza al menos 200GB MB de espacio para el consumo de snapshots.

Cuando `percent-snapshot-space` se establece en cero, todo el espacio del volumen está disponible para el usuario final, lo que proporciona una mejor visibilidad. Un administrador de bases de datos debe comprender que, si ve un volumen de 1TB GB que aprovecha las copias Snapshot, este espacio de 1TB TB se compartirá entre los datos activos y la rotación de copias Snapshot.

No hay una preferencia clara entre la opción uno y la opción dos entre los usuarios finales.

Snapshots de ONTAP y de terceros

El ID de documento de Oracle 604683,1 explica los requisitos para la compatibilidad con Snapshot de terceros y las múltiples opciones disponibles para las operaciones de backup y restauración.

El proveedor externo debe garantizar que las copias Snapshot de la empresa cumplen con los requisitos siguientes:

- Las copias Snapshot deben integrarse con las operaciones de restauración y recuperación recomendadas de Oracle.
- Las instantáneas deben ser consistentes con los fallos de la base de datos en el punto de la instantánea.
- El orden de escritura se conserva para cada archivo dentro de una instantánea.

Los productos de gestión de Oracle de ONTAP y NetApp cumplen estos requisitos.

SnapRestore

Restauración de datos rápida en ONTAP a partir de una copia Snapshot realizada por la tecnología NetApp SnapRestore.

Cuando un conjunto de datos críticos no está disponible, las operaciones empresariales fundamentales no funcionan. Las cintas pueden romperse e incluso las restauraciones de backups basados en discos pueden ser lentas para transferirse por la red. SnapRestore evita estos problemas al ofrecer una restauración casi instantánea de conjuntos de datos. Incluso las bases de datos con capacidad de petabytes se pueden restaurar por completo con tan solo unos minutos.

Hay dos formas de SnapRestore: Basado en archivos/LUN y basado en volúmenes.

- Pueden restaurarse archivos o LUN individuales en segundos, tanto si se trata de un LUN de 2TB GB como de un archivo 4KB.
- El contenedor de archivos o LUN se puede restaurar en segundos, ya sea 10GB o 100TB TB de datos.

Un «contenedor de archivos o LUN» normalmente hace referencia a un volumen FlexVol. Por ejemplo, puede tener 10 LUN que componen un grupo de discos LVM en un único volumen o un volumen puede almacenar los directorios iniciales NFS de 1000 usuarios. En lugar de ejecutar una operación de restauración para cada archivo o LUN individuales, puede restaurar el volumen completo como una única operación. Este proceso también funciona con contenedores de escalado horizontal que incluyen múltiples volúmenes, como una FlexGroup o un grupo de consistencia ONTAP.

La razón por la que SnapRestore funciona tan rápido y eficientemente se debe a la naturaleza de una copia Snapshot, que es esencialmente una vista paralela de solo lectura del contenido de un volumen en un momento determinado. Los bloques activos son los bloques reales que se pueden cambiar, mientras que la copia Snapshot es una vista de solo lectura del estado de los bloques que constituyen los archivos y la LUN en el momento de crear la copia Snapshot.

ONTAP solo permite el acceso de solo lectura a los datos de snapshots, pero los datos se pueden reactivar con SnapRestore. La copia de Snapshot se vuelve a habilitar como una vista de lectura y escritura de los datos, lo que devuelve los datos a su estado anterior. SnapRestore puede funcionar a nivel de volumen o archivo. La tecnología es esencialmente la misma con algunas pequeñas diferencias en el comportamiento.

SnapRestore de volumen

La SnapRestore basada en volúmenes devuelve todo el volumen de datos a un estado anterior. Esta operación no requiere el movimiento de datos, lo que significa que el proceso de restauración es esencialmente instantáneo, aunque la operación de la API o la CLI puede tardar unos segundos en procesarse. La restauración de 1GB TB de datos no es más complicada ni requiere más tiempo que restaurar 1PB TB de datos. Esta funcionalidad es el principal motivo por el que muchos clientes empresariales migran a los sistemas de almacenamiento de ONTAP. Proporciona un objetivo de tiempo de recuperación que se mide en segundos incluso para los conjuntos de datos de mayor tamaño.

Una desventaja de la SnapRestore basada en el volumen se debe al hecho de que los cambios dentro de un volumen son acumulativos con el tiempo. Por lo tanto, cada instantánea y los datos del archivo activo dependen de los cambios que conduzcan a ese punto. Revertir un volumen a un estado anterior implica descartar todos los cambios posteriores que se habían realizado en los datos. Sin embargo, lo que no resulta tan obvio es que se incluyen las instantáneas creadas posteriormente. Esto no siempre es deseable.

Por ejemplo, un acuerdo de nivel de servicio de retención de datos puede especificar 30 días de backups nocturnos. Si se restaura un conjunto de datos en una snapshot creada hace cinco días con SnapRestore para volúmenes, se descartarán todas las snapshots creadas en los cinco días anteriores, lo que infringe el acuerdo de nivel de servicio.

Hay varias opciones disponibles para abordar esta limitación:

1. Los datos se pueden copiar a partir de una snapshot anterior, en lugar de realizar una SnapRestore de todo el volumen. Este método funciona mejor con conjuntos de datos más pequeños.
2. Una copia Snapshot puede clonarse en lugar de restaurarse. La limitación de este enfoque es que la copia Snapshot de origen depende del clon. Por lo tanto, no se puede eliminar a menos que también se elimine el clon o se divida en un volumen independiente.
3. Uso de SnapRestore basado en archivos.

SnapRestore de archivos

La SnapRestore basada en archivos es un proceso de restauración más granular basado en Snapshot. En lugar de revertir el estado de un volumen completo, se revierte el estado de un archivo individual o LUN. No es necesario eliminar ninguna instantánea, ni esta operación crea ninguna dependencia de una instantánea anterior. El archivo o el LUN estarán disponibles de inmediato en el volumen activo.

No es necesario mover datos durante una restauración SnapRestore de un archivo o una LUN. Sin embargo, se requieren algunas actualizaciones internas de metadatos para reflejar el hecho de que los bloques subyacentes de un archivo o LUN ahora existen tanto en una snapshot como en el volumen activo. No debería afectar el rendimiento, pero este proceso bloquea la creación de snapshots hasta que se completa. La tasa de procesamiento es de aproximadamente 5Gbps (18TB TB/hora) en función del tamaño total de los archivos restaurados.

Backups en línea

Se necesitan dos conjuntos de datos para proteger y recuperar una base de datos de Oracle en modo de backup. Tenga en cuenta que esta no es la única opción de copia de seguridad de Oracle, pero es la más común.

- Instantánea de los archivos de datos en modo de copia de seguridad
- Los registros de archivos creados mientras los archivos de datos estaban en modo de backup

Si se necesita una recuperación completa, incluidas todas las transacciones confirmadas, se requiere un tercer elemento:

- Juego de redo logs actuales

Existen varias formas de impulsar la recuperación de un backup en línea. Muchos clientes restauran snapshots mediante la interfaz de línea de comandos de ONTAP y, a continuación, usando Oracle RMAN o sqlplus para completar la recuperación. Esto es especialmente habitual en entornos de producción de gran tamaño en los que la probabilidad y frecuencia de las restauraciones de bases de datos es extremadamente baja y cualquier procedimiento de restauración lo gestiona un administrador de bases de datos cualificado. Para obtener una automatización completa, las soluciones como NetApp SnapCenter incluyen un complemento de Oracle con interfaces gráficas y de línea de comandos.

Algunos clientes a gran escala han adoptado un enfoque más simple mediante la configuración de secuencias de comandos básicas en los hosts para colocar las bases de datos en modo de backup en un momento específico de preparación para una copia Snapshot programada. Por ejemplo, programe el comando `alter database begin backup` a las 23:58, `alter database end backup` a las 00:02, y después programe copias snapshot directamente en el sistema de almacenamiento a medianoche. El resultado es una estrategia de backup sencilla y altamente escalable que no requiere software ni licencias externas.

Distribución de datos

El diseño más sencillo es aislar los archivos de datos en uno o varios volúmenes dedicados. No deben estar contaminados por ningún otro tipo de archivo. De este modo, se garantiza que los volúmenes de archivos de datos puedan restaurarse rápidamente mediante una operación SnapRestore sin destruir un registro de recuperación, un archivo de control o un archivo importante.

SAN tiene requisitos similares para aislamiento de archivos de datos en volúmenes dedicados. Con un sistema operativo como Microsoft Windows, un único volumen puede contener varios LUN de archivos de datos, cada uno con un sistema de archivos NTFS. Con otros sistemas operativos, generalmente hay un administrador de volúmenes lógicos. Por ejemplo, con Oracle ASM, la opción más sencilla sería confinar los LUN de un grupo de discos ASM en un único volumen del que se pueda incluir y restaurar como unidad en un backup. Si se necesitan volúmenes adicionales por motivos de rendimiento o gestión de capacidad, crear un grupo de discos adicional en el nuevo volumen simplifica la gestión.

Si se siguen estas directrices, se pueden programar Snapshot directamente en el sistema de almacenamiento sin requisitos para realizar una snapshot de grupo de coherencia. El motivo es que las copias de seguridad de Oracle no necesitan que se realice una copia de seguridad de los archivos de datos al mismo tiempo. El procedimiento de backup online se diseñó para permitir que los archivos de datos sigan actualizándose a medida que se transmiten lentamente a la cinta durante horas.

Se produce una complicación en situaciones como el uso de un grupo de discos de ASM que se distribuye entre volúmenes. En estos casos, se debe realizar una cg-snapshot para garantizar que los metadatos de ASM sean coherentes en todos los volúmenes constituyentes.

Precaución: Verifique que el ASM `spfile` y.. `passwd` los archivos no están en el grupo de discos que aloja los archivos de datos. Esto interfiere con la capacidad de restaurar selectivamente archivos de datos y solo archivos de datos.

Procedimiento de recuperación local: NFS

Este procedimiento se puede realizar manualmente o a través de una aplicación como SnapCenter. El procedimiento básico es el siguiente:

1. Cierre la base de datos.

2. Recupere los volúmenes del archivo de datos en la instantánea inmediatamente antes del punto de restauración deseado.
3. Reproduzca los archive logs en el punto deseado.
4. Reproduzca los redo logs actuales si desea una recuperación completa.

En este procedimiento se asume que los archive logs deseados siguen presentes en el sistema de archivos activo. De lo contrario, se deben restaurar los archive logs o se puede dirigir rman/sqlplus a los datos del directorio de instantáneas.

Además, para bases de datos más pequeñas, un usuario final puede recuperar archivos de datos directamente desde `.snapshot` directorio sin la ayuda de herramientas de automatización o administradores del almacenamiento para ejecutar un `snaprestore` comando.

Procedimiento de recuperación local: San

Este procedimiento se puede realizar manualmente o a través de una aplicación como SnapCenter. El procedimiento básico es el siguiente:

1. Cierre la base de datos.
2. Desactive los grupos de discos que alojan los archivos de datos. El procedimiento varía en función del gestor de volúmenes lógico elegido. Con ASM, el proceso requiere desmontar el grupo de discos. Con Linux, los sistemas de archivos deben desmontarse y los volúmenes lógicos y los grupos de volúmenes deben desactivarse. El objetivo es detener todas las actualizaciones en el grupo de volúmenes objetivo que se va a restaurar.
3. Restaure los grupos de discos de archivos de datos en la instantánea inmediatamente antes del punto de restauración deseado.
4. Vuelva a activar los grupos de discos recién restaurados.
5. Reproduzca los archive logs en el punto deseado.
6. Vuelva a reproducir todos los redo logs si desea realizar una recuperación completa.

En este procedimiento se asume que los archive logs deseados siguen presentes en el sistema de archivos activo. Si no lo son, los registros de archivos se deben restaurar desconectando las LUN del registro de archivos y ejecutando una restauración. Este es también un ejemplo en el que la división de archive logs en volúmenes dedicados es útil. Si los registros de archivos comparten un grupo de volúmenes con registros de recuperación, se deben copiar en otro lugar los registros de recuperación antes de restaurar el conjunto general de LUN. Este paso evita la pérdida de las transacciones registradas finales.

Backups optimizados para Snapshot de almacenamiento

Cuando se lanzó Oracle 12c, ya que no es necesario colocar una base de datos en modo de backup dinámico, se simplificaron aún más las tareas de backup y recuperación basadas en Snapshots. El resultado es la capacidad de programar backups basados en snapshots directamente en un sistema de almacenamiento y mantener la capacidad para realizar una recuperación completa o de un momento específico.

Aunque el procedimiento de recuperación de backup dinámico es más familiar para los administradores de bases de datos, durante mucho tiempo ha sido posible usar snapshots que no se crearon mientras la base de datos estaba en modo de backup dinámico. Oracle 10g y 11g requerían pasos manuales adicionales durante la recuperación para hacer que la base de datos fuera coherente. Con Oracle 12c, `sqlplus` y.. `rman` contienen la lógica adicional para reproducir archive logs en copias de seguridad de archivos de datos que no

estaban en modo de copia de seguridad activa.

Como hemos visto anteriormente, la recuperación de un backup en caliente basado en instantáneas requiere dos conjuntos de datos:

- Instantánea de los archivos de datos creados en modo de backup
- Los registros de archivos generados mientras los archivos de datos estaban en modo de backup dinámico

Durante la recuperación, la base de datos lee los metadatos de los archivos de datos para seleccionar los archive logs requeridos para la recuperación.

La recuperación optimizada para snapshot de almacenamiento requiere conjuntos de datos ligeramente diferentes para lograr los mismos resultados:

- Una instantánea de los archivos de datos, además de un método para identificar la hora a la que se creó la instantánea
- Archive logs desde la hora del punto de control del archivo de datos más reciente hasta la hora exacta de la instantánea

Durante la recuperación, la base de datos lee metadatos de los archivos de datos para identificar el primer archive log necesario. Se puede realizar una recuperación completa o a un momento específico. Al realizar una recuperación puntual, es fundamental conocer la hora de la instantánea de los archivos de datos. El punto de recuperación especificado debe ser posterior a la hora de creación de las instantáneas. NetApp recomienda añadir al menos unos minutos al tiempo de la snapshot para justificar la variación de reloj.

Para obtener más información, consulte la documentación de Oracle sobre el tema «Recuperación mediante la optimización de instantáneas de almacenamiento» disponible en varias versiones de la documentación de Oracle 12c. Además, consulte el ID de documento de Oracle 604683,1 con respecto al soporte de instantáneas de terceros de Oracle.

Distribución de datos

El diseño más sencillo es aislar los archivos de datos en uno o varios volúmenes dedicados. No deben estar contaminados por ningún otro tipo de archivo. De este modo, se garantiza que los volúmenes de archivos de datos se puedan restaurar rápidamente con una operación de SnapRestore sin destruir un registro de recuperación, un archivo de control o un archivo importante.

SAN tiene requisitos similares para aislamiento de archivos de datos en volúmenes dedicados. Con un sistema operativo como Microsoft Windows, un único volumen puede contener varios LUN de archivos de datos, cada uno con un sistema de archivos NTFS. Con otros sistemas operativos, generalmente hay un gestor de volúmenes lógicos también. Por ejemplo, con Oracle ASM, la opción más sencilla sería restringir los grupos de discos en un único volumen del que se pueda realizar un backup y restaurar como unidad. Si se necesitan volúmenes adicionales por motivos de rendimiento o gestión de capacidad, crear un grupo de discos adicional en el nuevo volumen simplifica la gestión.

Si se siguen estas directrices, se pueden programar Snapshot directamente en ONTAP sin requisitos para realizar una snapshot de grupo de coherencia. El motivo es que las copias de seguridad optimizadas para instantáneas no necesitan que se realice una copia de seguridad de los archivos de datos al mismo tiempo.

Se produce una complicación en situaciones como un grupo de discos de ASM que se distribuye entre volúmenes. En estos casos, se debe realizar una cg-snapshot para garantizar que los metadatos de ASM sean coherentes en todos los volúmenes constituyentes.

[Nota]Verifique que los archivos spfile y passwd de ASM no estén en el grupo de discos que aloja los archivos

de datos. Esto interfiere con la capacidad de restaurar selectivamente archivos de datos y solo archivos de datos.

Procedimiento de recuperación local: NFS

Este procedimiento se puede realizar manualmente o a través de una aplicación como SnapCenter. El procedimiento básico es el siguiente:

1. Cierre la base de datos.
2. Recupere los volúmenes del archivo de datos en la instantánea inmediatamente antes del punto de restauración deseado.
3. Reproduzca los archive logs en el punto deseado.

En este procedimiento se asume que los archive logs deseados siguen presentes en el sistema de archivos activo. Si no lo son, se deben restaurar los registros de archivos `rman` o. `sqlplus` se puede dirigir a los datos de la `.snapshot` directorio.

Además, para bases de datos más pequeñas, un usuario final puede recuperar archivos de datos directamente desde `.snapshot` Directorio sin ayuda de las herramientas de automatización o de un administrador del almacenamiento para ejecutar un comando de la SnapRestore.

Procedimiento de recuperación local: San

Este procedimiento se puede realizar manualmente o a través de una aplicación como SnapCenter. El procedimiento básico es el siguiente:

1. Cierre la base de datos.
2. Desactive los grupos de discos que alojan los archivos de datos. El procedimiento varía en función del gestor de volúmenes lógico elegido. Con ASM, el proceso requiere desmontar el grupo de discos. Con Linux, los sistemas de archivos deben desmontarse y los volúmenes lógicos y los grupos de volúmenes están desactivados. El objetivo es detener todas las actualizaciones en el grupo de volúmenes objetivo que se va a restaurar.
3. Restaure los grupos de discos de archivos de datos en la instantánea inmediatamente antes del punto de restauración deseado.
4. Vuelva a activar los grupos de discos recién restaurados.
5. Reproduzca los archive logs en el punto deseado.

En este procedimiento se asume que los archive logs deseados siguen presentes en el sistema de archivos activo. Si no lo son, los registros de archivos se deben restaurar desconectando las LUN del registro de archivos y ejecutando una restauración. Este es también un ejemplo en el que la división de archive logs en volúmenes dedicados es útil. Si los registros de archivos comparten un grupo de volúmenes con redo logs, los redo logs se deben copiar en otro lugar antes de restaurar el conjunto general de LUN para evitar perder las transacciones finales registradas.

Ejemplo de recuperación completa

Supongamos que los archivos de datos se han dañado o destruido y se necesita una recuperación completa. El procedimiento para hacerlo es el siguiente:

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover automatic;
Media recovery complete.
SQL> alter database open;
Database altered.
SQL>
```

Ejemplo de recuperación a un momento específico

Todo el procedimiento de recuperación es un único comando: `recover automatic`.

Si se requiere una recuperación a un momento específico, es necesario conocer la marca de hora de las instantáneas y se puede identificar de la siguiente manera:

```
Cluster01::> snapshot show -vserver vserver1 -volume NTAP_oradata -fields
create-time
vserver   volume        snapshot   create-time
--------  ------------  ---------  ------------------------
vserver1  NTAP_oradata  my-backup  Thu Mar 09 10:10:06 2017
```

La hora de creación de la copia Snapshot se muestra como 9th de marzo y 10:10:06. Para estar seguro, se añade un minuto a la hora de la copia Snapshot:

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover database until time '09-MAR-2017 10:44:15' snapshot time '09-
MAR-2017 10:11:00';
```

La recuperación se inicia ahora. Especificó una hora de instantánea de 10:11:00, un minuto después del tiempo registrado para contabilizar la posible variación de reloj y un tiempo de recuperación objetivo de 10:44. A continuación, sqlplus solicita los archive logs necesarios para alcanzar el tiempo de recuperación deseado de 10:44.

```
ORA-00279: change 551760 generated at 03/09/2017 05:06:07 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_31_930813377.dbf
ORA-00280: change 551760 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 552566 generated at 03/09/2017 05:08:09 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_32_930813377.dbf
ORA-00280: change 552566 for thread 1 is in sequence #32
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 553045 generated at 03/09/2017 05:10:12 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_33_930813377.dbf
ORA-00280: change 553045 for thread 1 is in sequence #33
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 753229 generated at 03/09/2017 05:15:58 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_34_930813377.dbf
ORA-00280: change 753229 for thread 1 is in sequence #34
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
Log applied.
Media recovery complete.
SQL> alter database open resetlogs;
Database altered.
SQL>
```

Recuperación completa de una base de datos utilizando instantáneas utilizando el `recover automatic` el comando no requiere una licencia específica, sino un uso de recuperación puntual `snapshot time` Necesita la licencia de Oracle Advanced Compression.

Herramientas de automatización y gestión de base de datos

El valor principal de ONTAP en un entorno de bases de datos de Oracle proviene de las tecnologías principales de ONTAP, como las copias Snapshot instantáneas, la replicación simple de SnapMirror y la creación eficiente de los volúmenes FlexClone.

En algunos casos, la simple configuración de estas funciones básicas directamente en ONTAP satisface los requisitos, pero las necesidades más complicadas requieren una capa de orquestación.

SnapCenter

SnapCenter es el producto estrella de protección de datos de NetApp. A un nivel muy bajo, es similar a los

productos de SnapManager en cuanto a cómo se ejecutan backups de bases de datos, pero se creó desde cero para proporcionar un panel único para la gestión de la protección de datos en sistemas de almacenamiento de NetApp.

SnapCenter incluye las funciones básicas, como los backups y restauraciones basados en Snapshot, la replicación de SnapMirror y SnapVault, y otras funciones necesarias para funcionar a escala para grandes empresas. Estas funciones avanzadas incluyen una funcionalidad ampliada de control de acceso basado en roles (RBAC), API RESTful para integrarse con productos de orquestación de terceros, gestión central no disruptiva de complementos de SnapCenter en hosts de bases de datos y una interfaz de usuario diseñada para entornos a escala de cloud.

DESCANSO

ONTAP también contiene un amplio conjunto de API RESTful. Esto permite que 3rd proveedores de partes creen protección de datos y otras aplicaciones de gestión con la profunda integración con ONTAP. Además, los clientes que desean crear sus propios flujos de trabajo y utilidades de automatización pueden consumir fácilmente la API RESTful.

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.

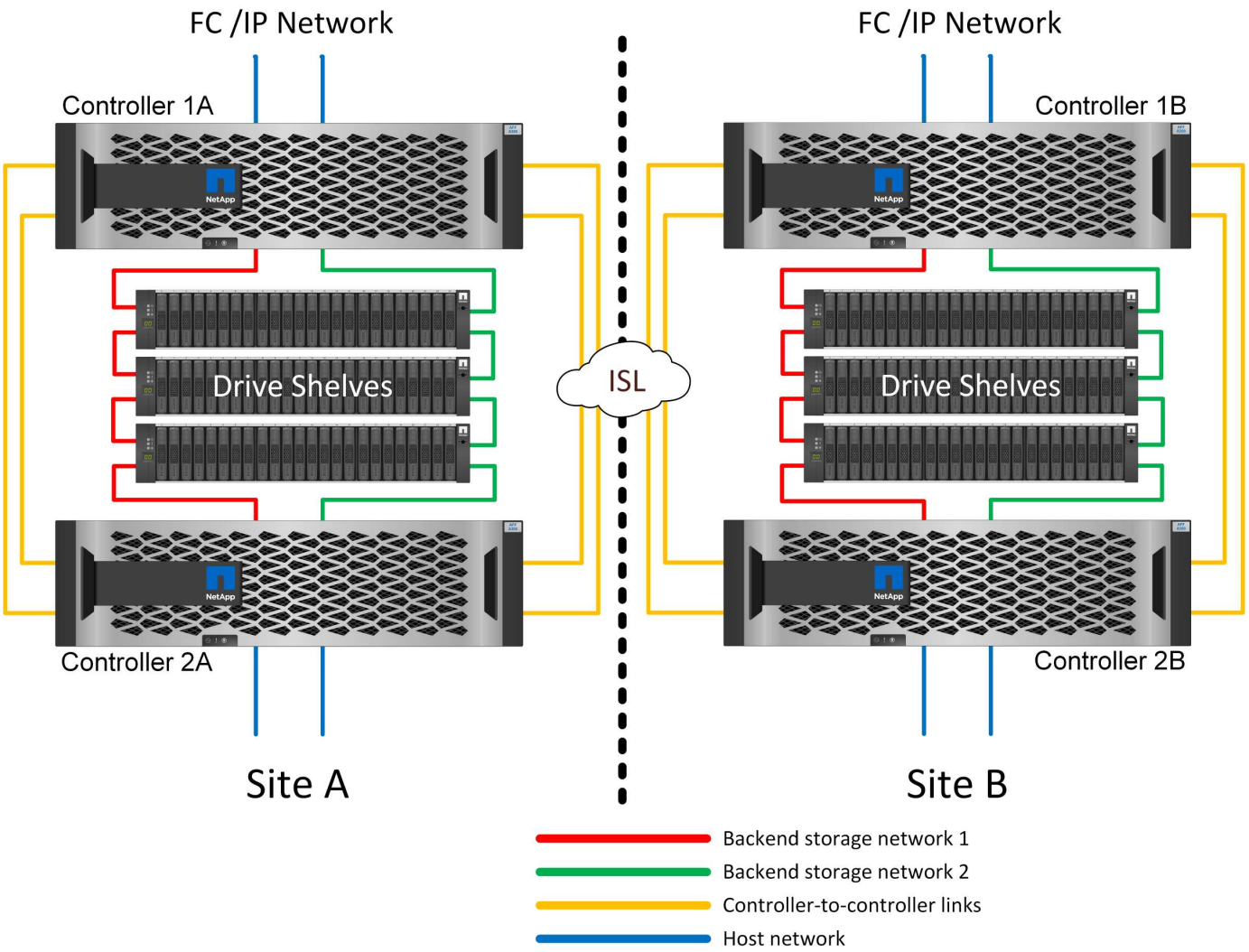
IP de MetroCluster

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.

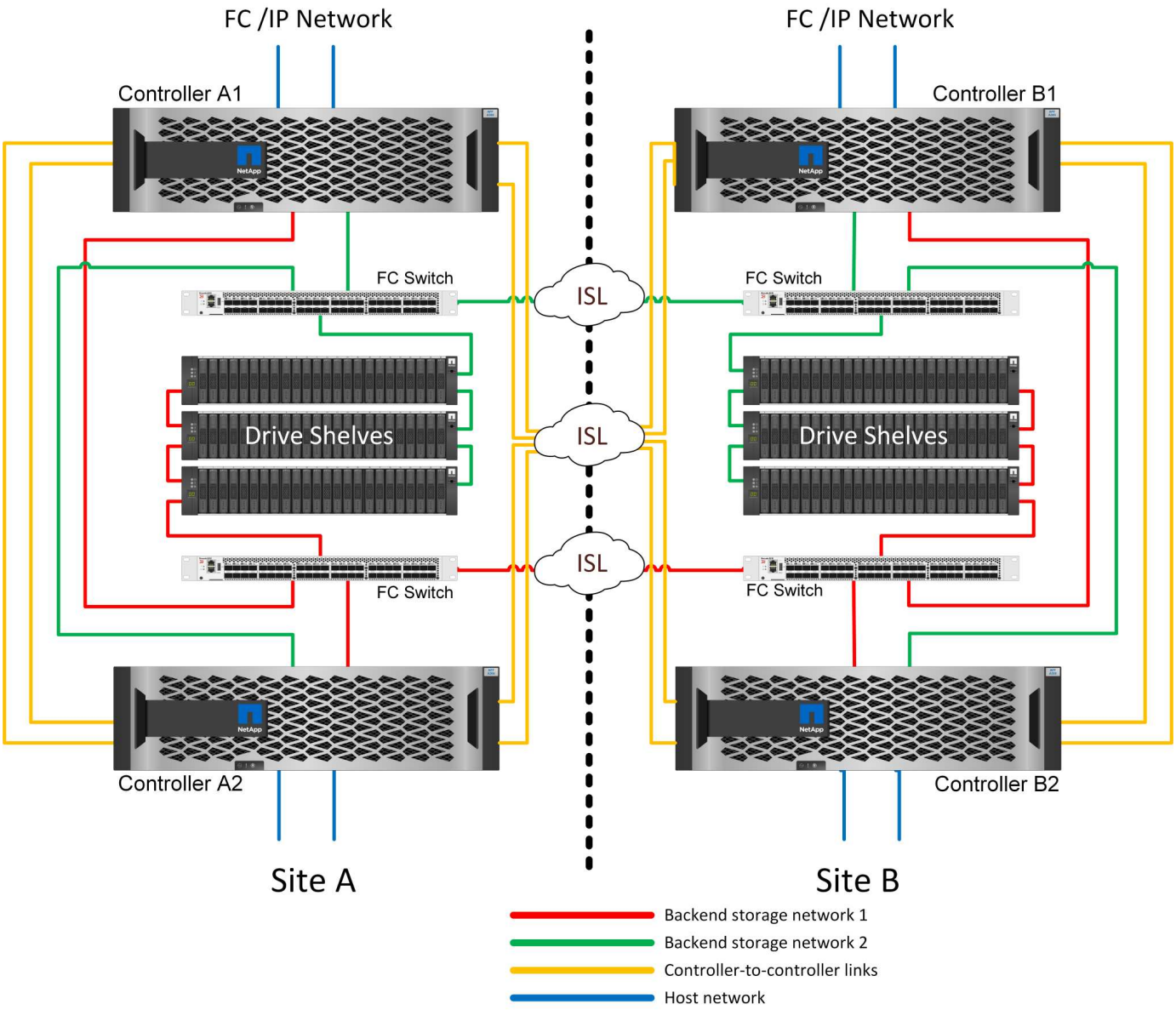
La arquitectura además es más sencilla ya que las únicas conexiones entre sitios son para las controladoras. En MetroCluster conectados a FC SAN, una controladora escribe directamente en las unidades del sitio opuesto y, por lo tanto, requiere conexiones SAN, switches y puentes adicionales. En cambio,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



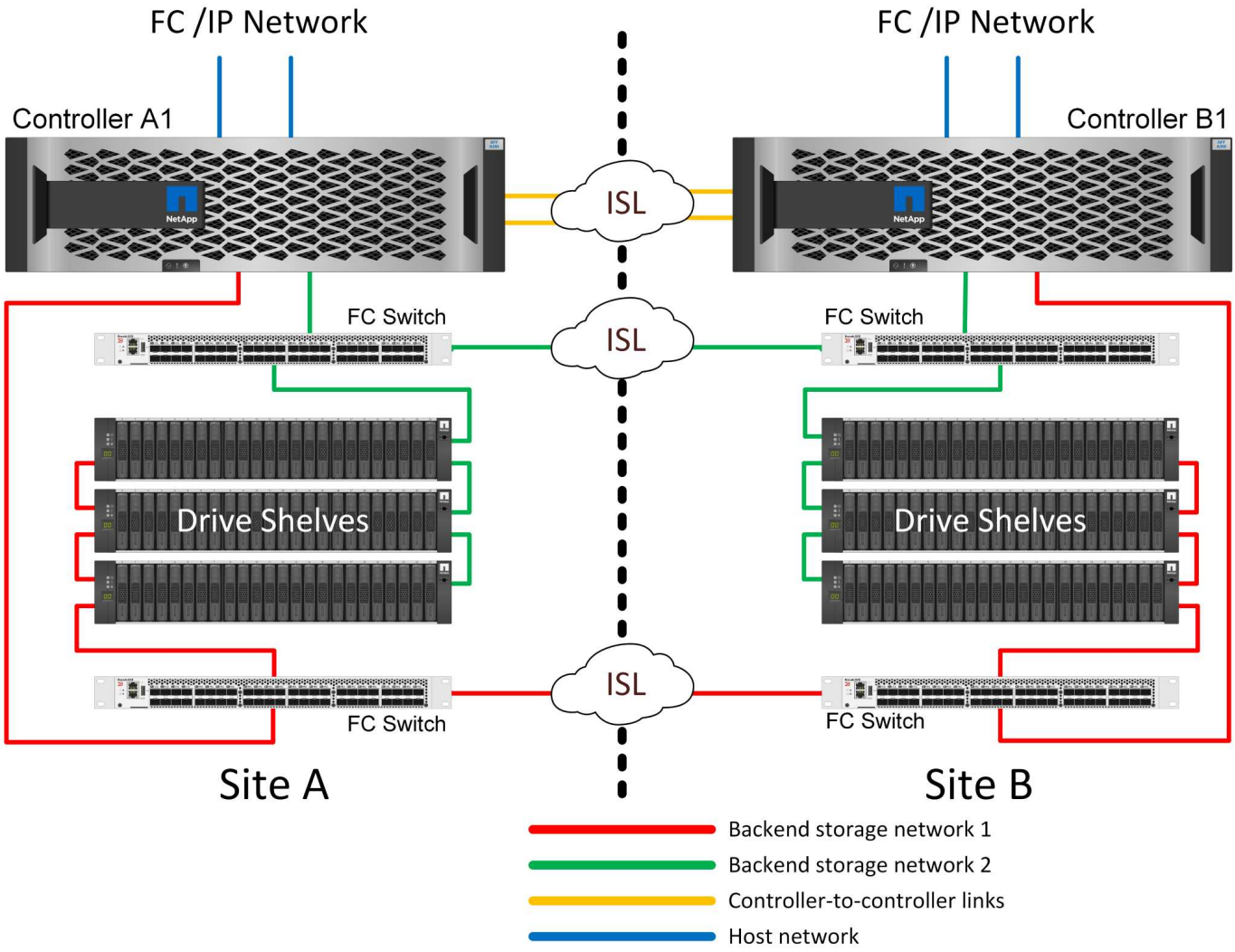
Algunas infraestructuras multisitio no están diseñadas para operaciones activo-activo, sino que se utilizan más como sitio principal y sitio de recuperación de desastres. En esta situación, generalmente es preferible una opción MetroCluster de una pareja de alta disponibilidad por las siguientes razones:

- 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

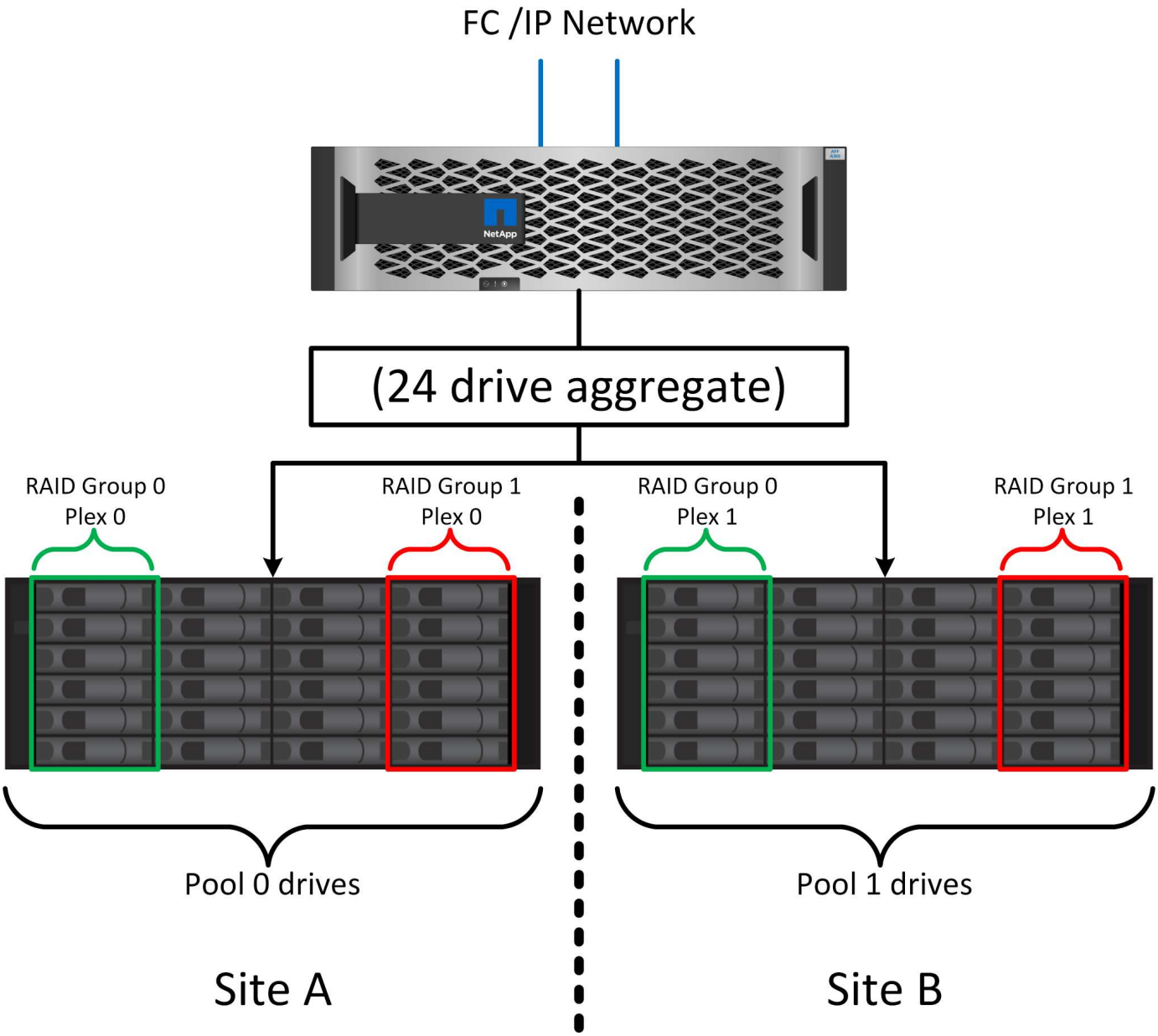
MetroCluster amplía la protección de datos de NVRAM de las siguientes formas:

- 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.

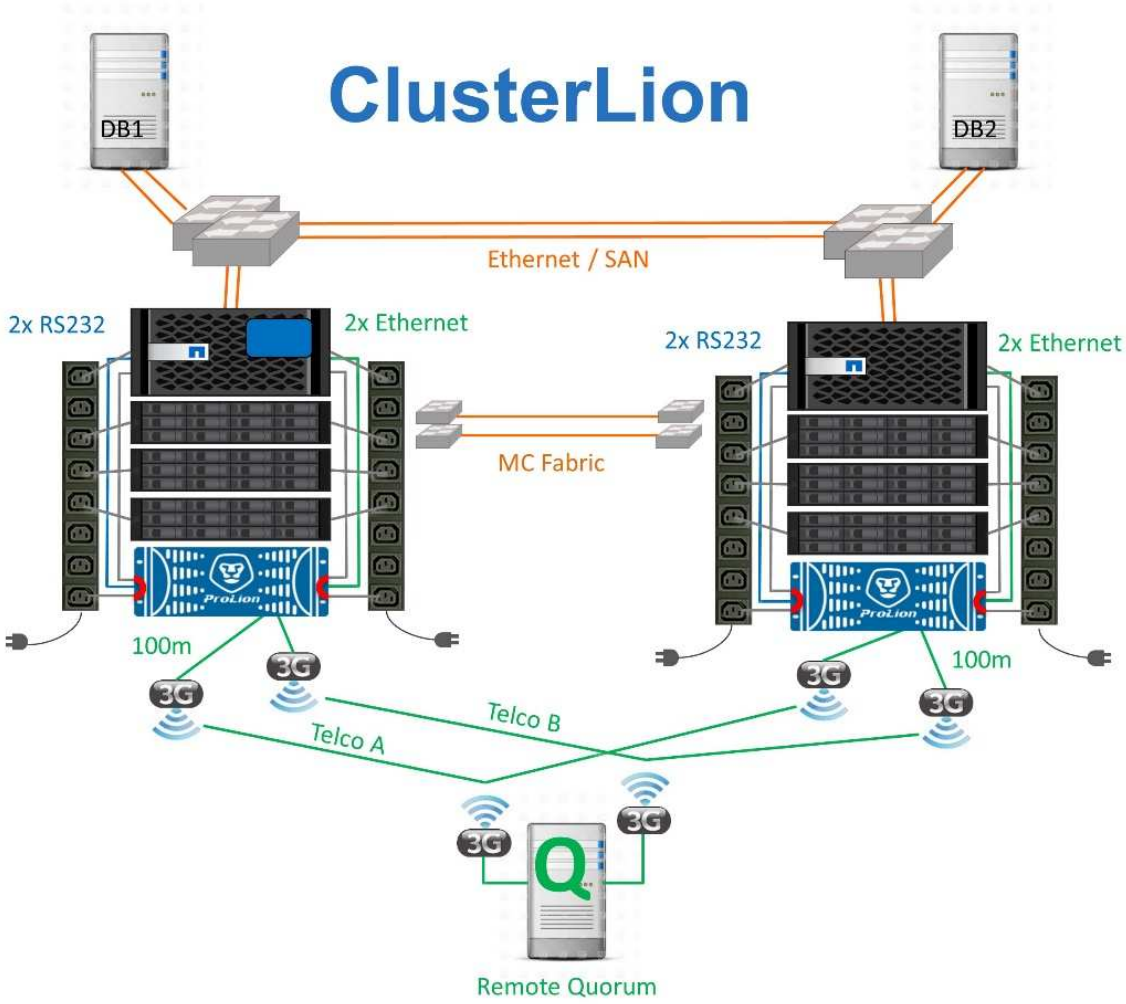
Una MetroCluster conectada a estructura tiene acceso directo a dispositivos de almacenamiento en el sitio opuesto. Esto permite que una controladora MetroCluster supervise el estado de las otras controladoras mediante la lectura de datos de latidos de las unidades. Esto permite que una controladora reconozca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

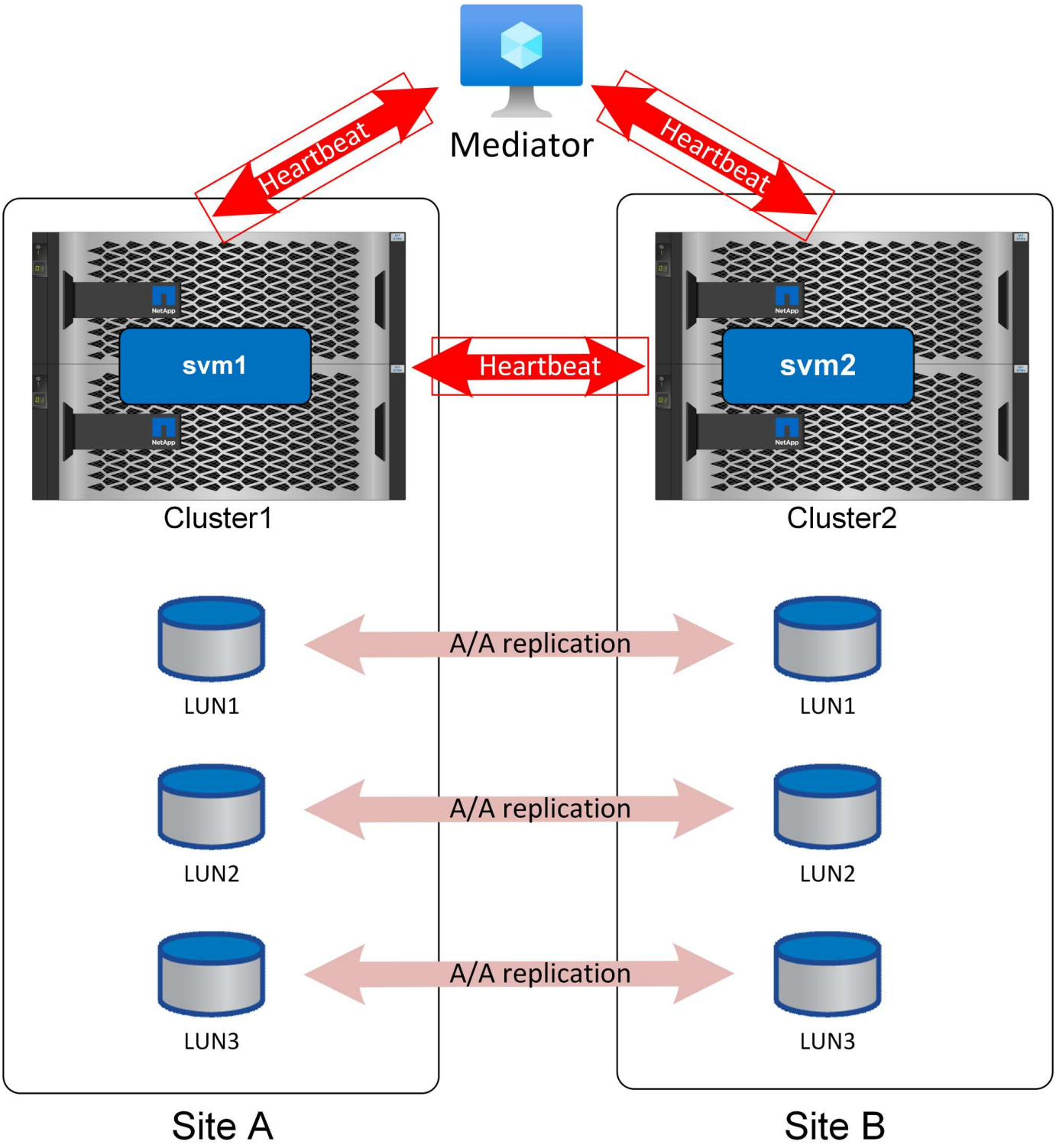
ONTAP Mediator es una aplicación de software que se descarga del soporte técnico de NetApp y que normalmente se implementa en una pequeña máquina virtual. ONTAP Mediator no es un tiebreaker cuando se utiliza con SnapMirror sincronización activa. Es un canal de comunicación alternativo para los dos clústeres 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.

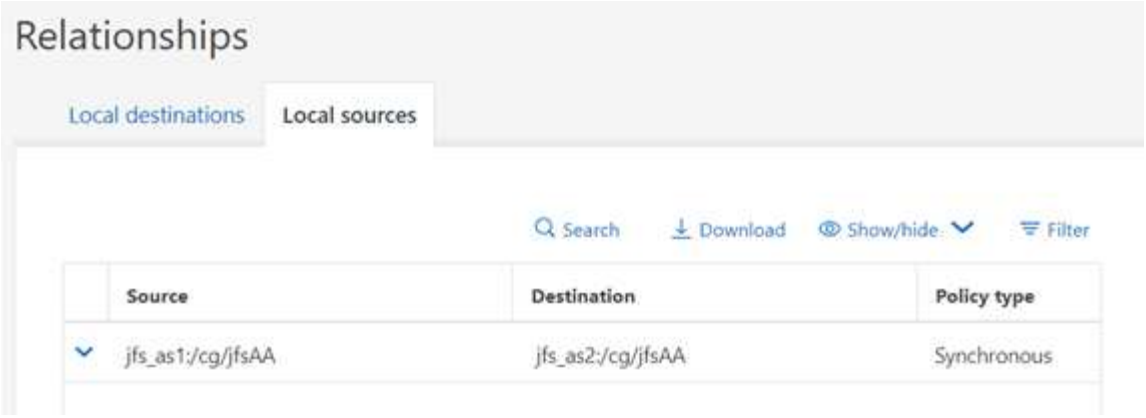
Sitio preferido de sincronización activa de SnapMirror

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.

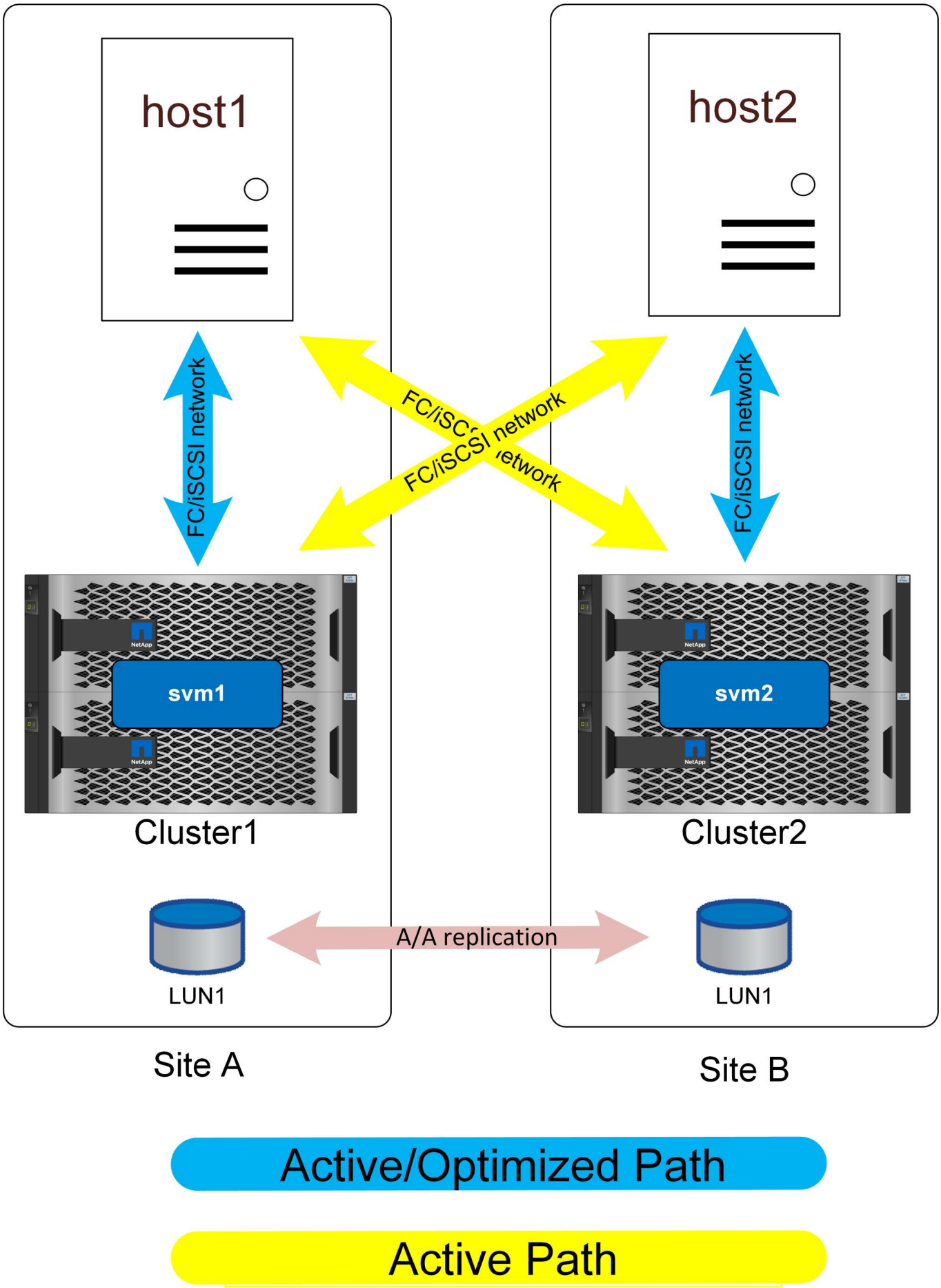
El segundo paso opcional es configurar la proximidad del host. Esto no controla el acceso, controla *priority*.

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

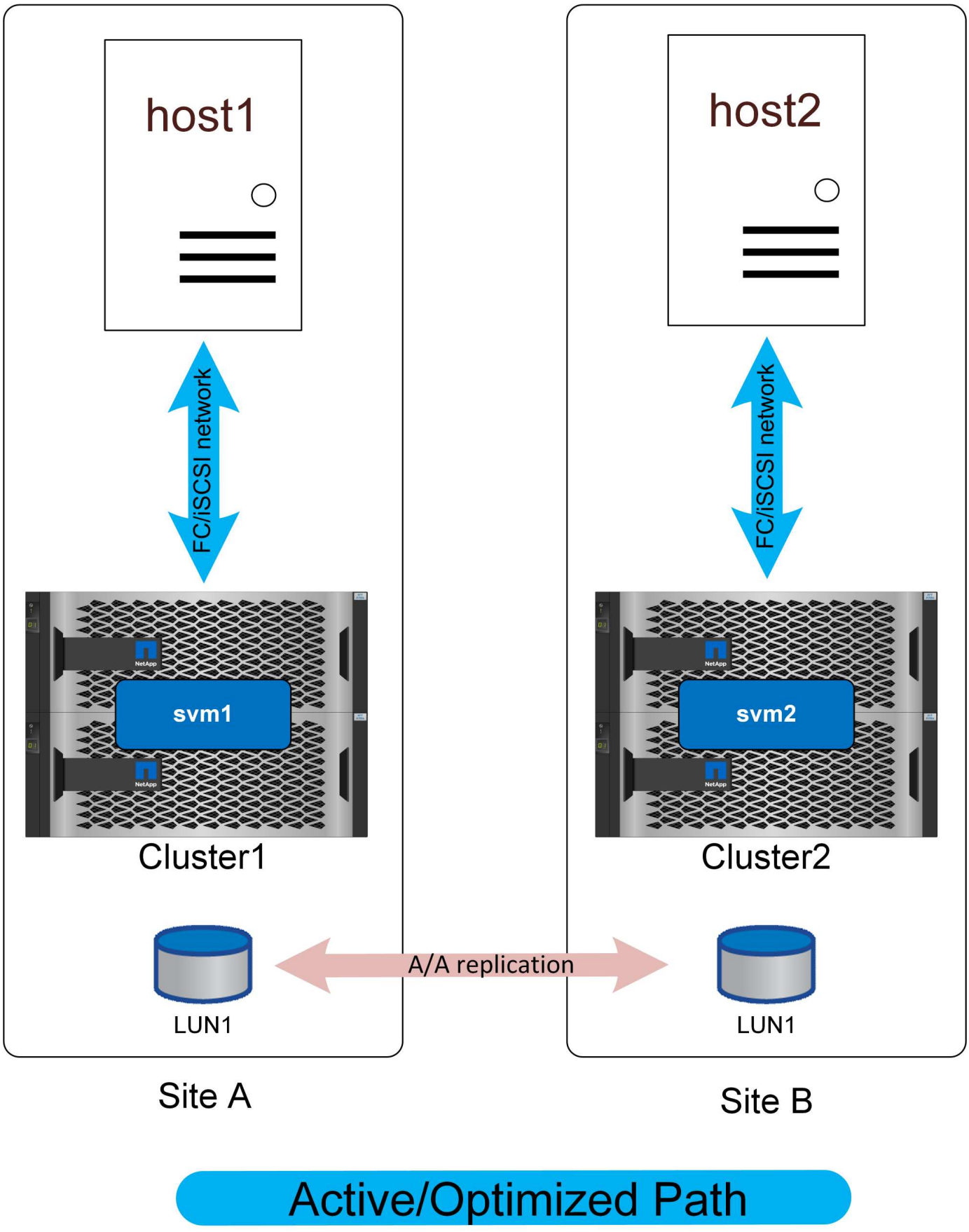
En funcionamiento normal, todas las I/O son locales. Las lecturas y escrituras se sirven desde la cabina de almacenamiento local. Por supuesto, el controlador local también deberá replicar el I/O de escritura en el sistema remoto antes de reconocerlo, pero todas las I/O de lectura se prestarán de servicio 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.



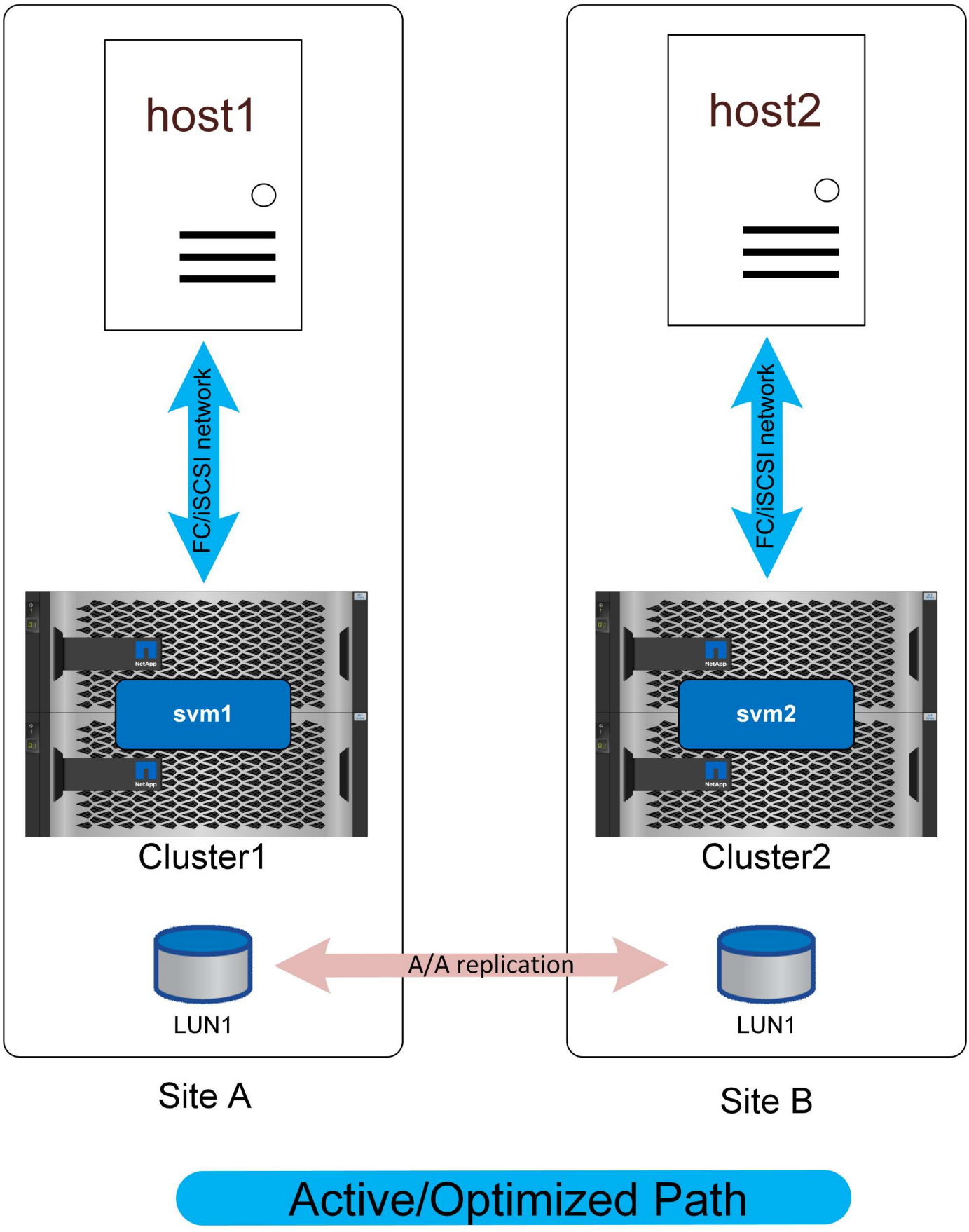
Una configuración de ASA con acceso no uniforme funcionaría en gran medida igual que con AFF. Con un acceso uniforme, IO estaría cruzando la WAN. Esto puede o no ser deseable.

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.

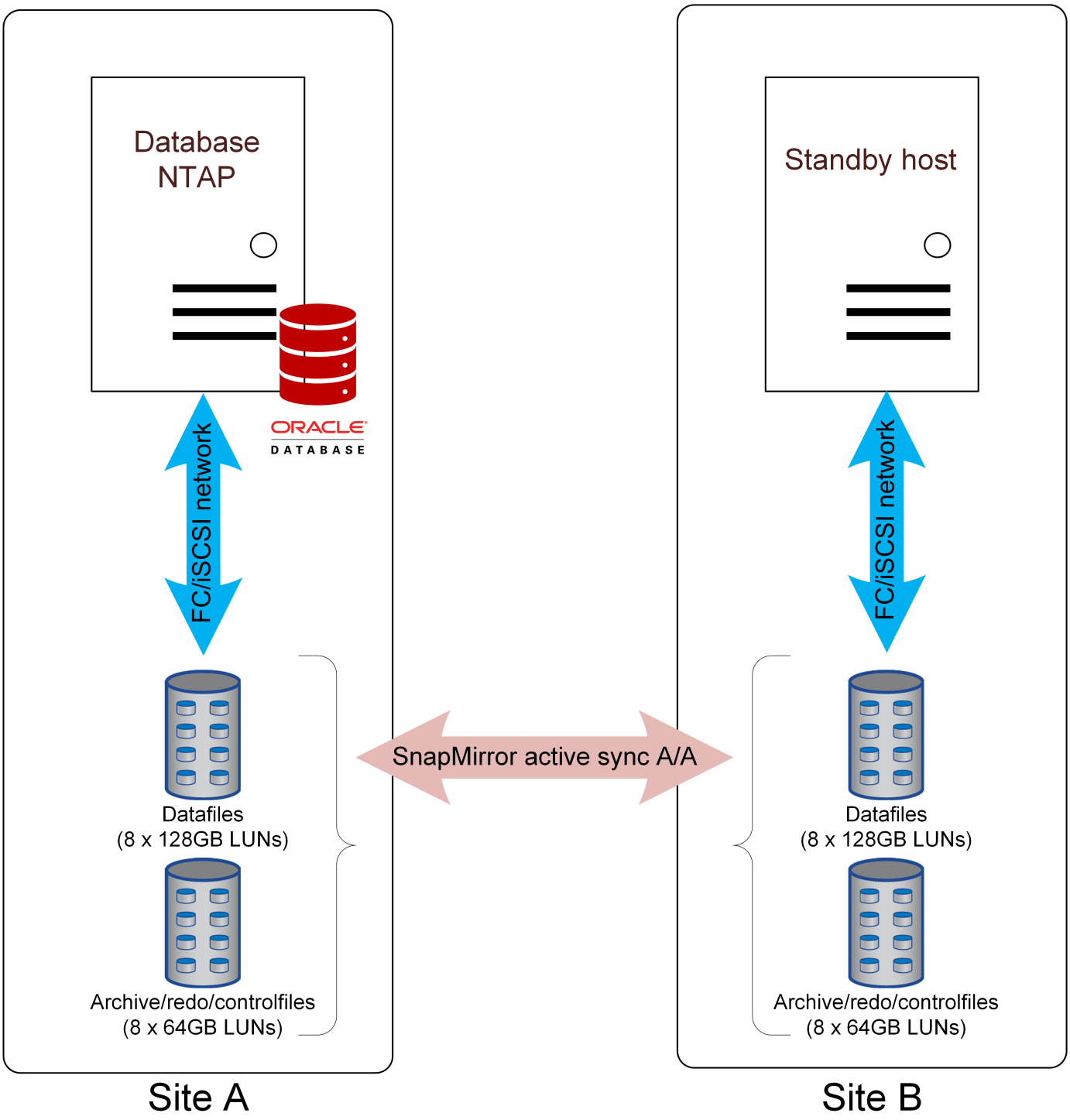
Si el objetivo de tiempo de recuperación fuera más estricto, la automatización activo-pasiva básica mediante scripts o clusterware como Pacemaker o Ansible mejoraría el tiempo de conmutación al nodo de respaldo. Por ejemplo, VMware HA podría configurarse para detectar un fallo de los equipos virtuales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

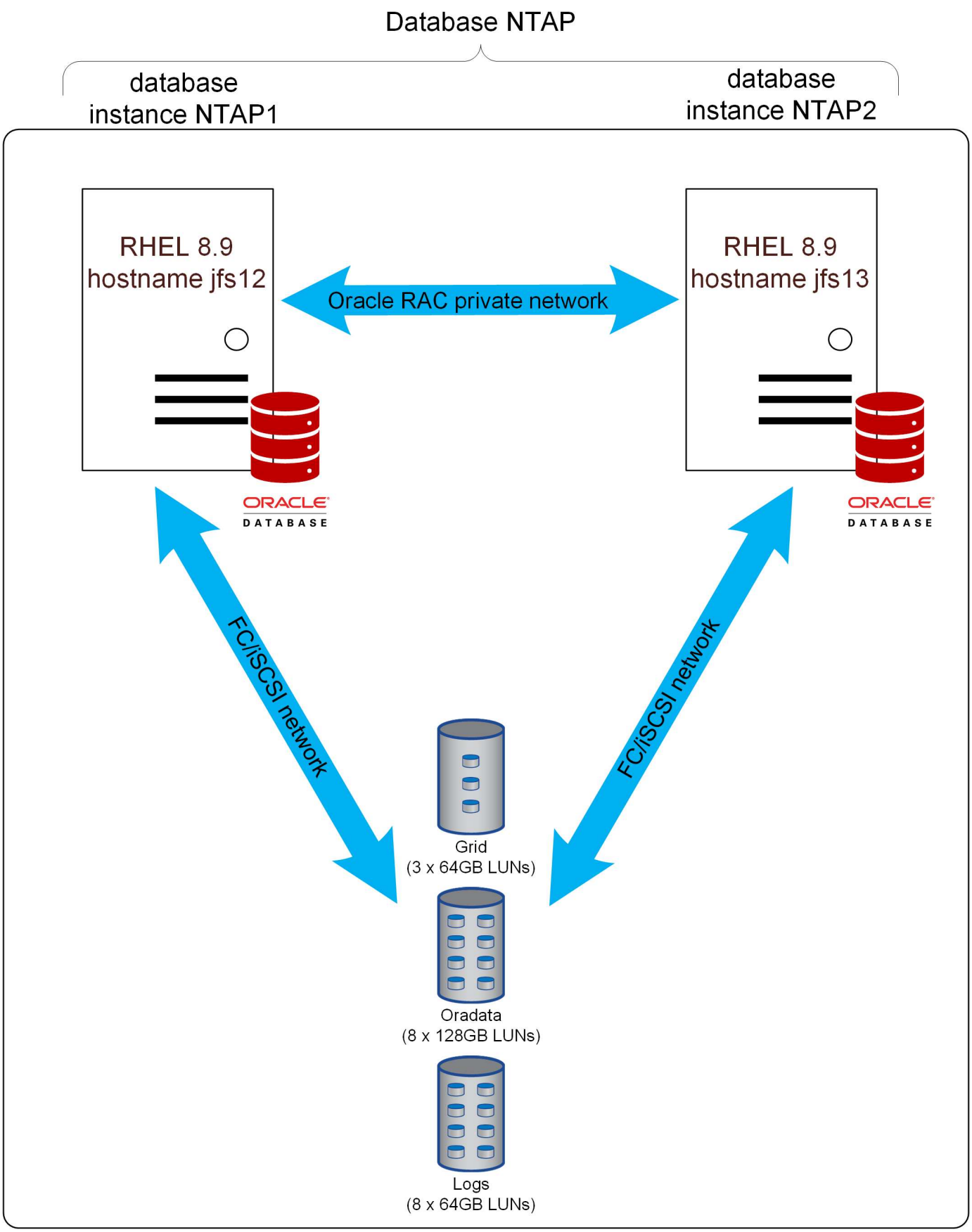
Oracle Extended RAC

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

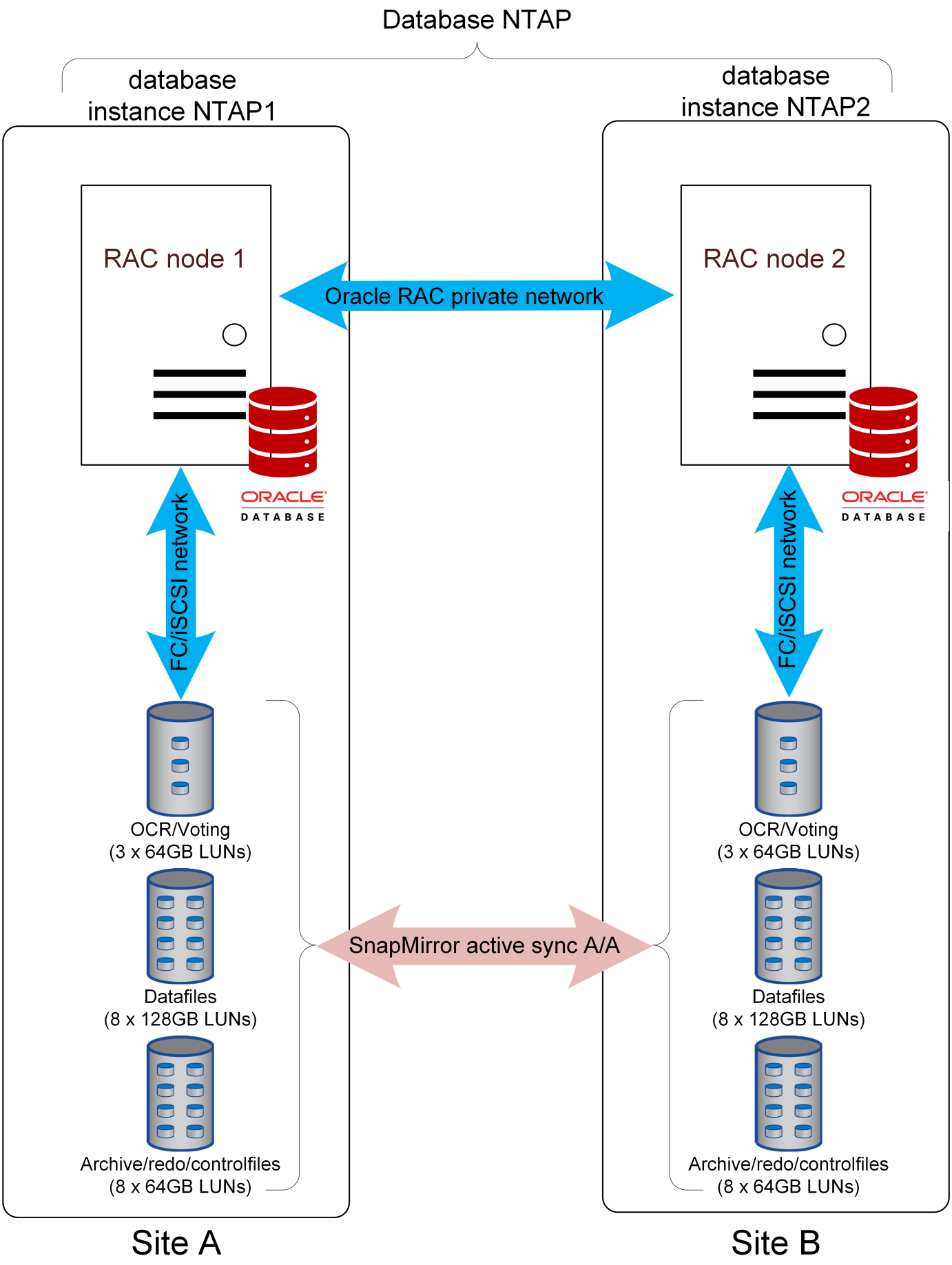
La consideración más importante con Oracle RAC en la sincronización activa de SnapMirror es si se debe utilizar un acceso uniforme o no uniforme.

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
ORACLE DATABASE
ORACLE DATABASE
FC/iSCSI network
FC/iSCSI network
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

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

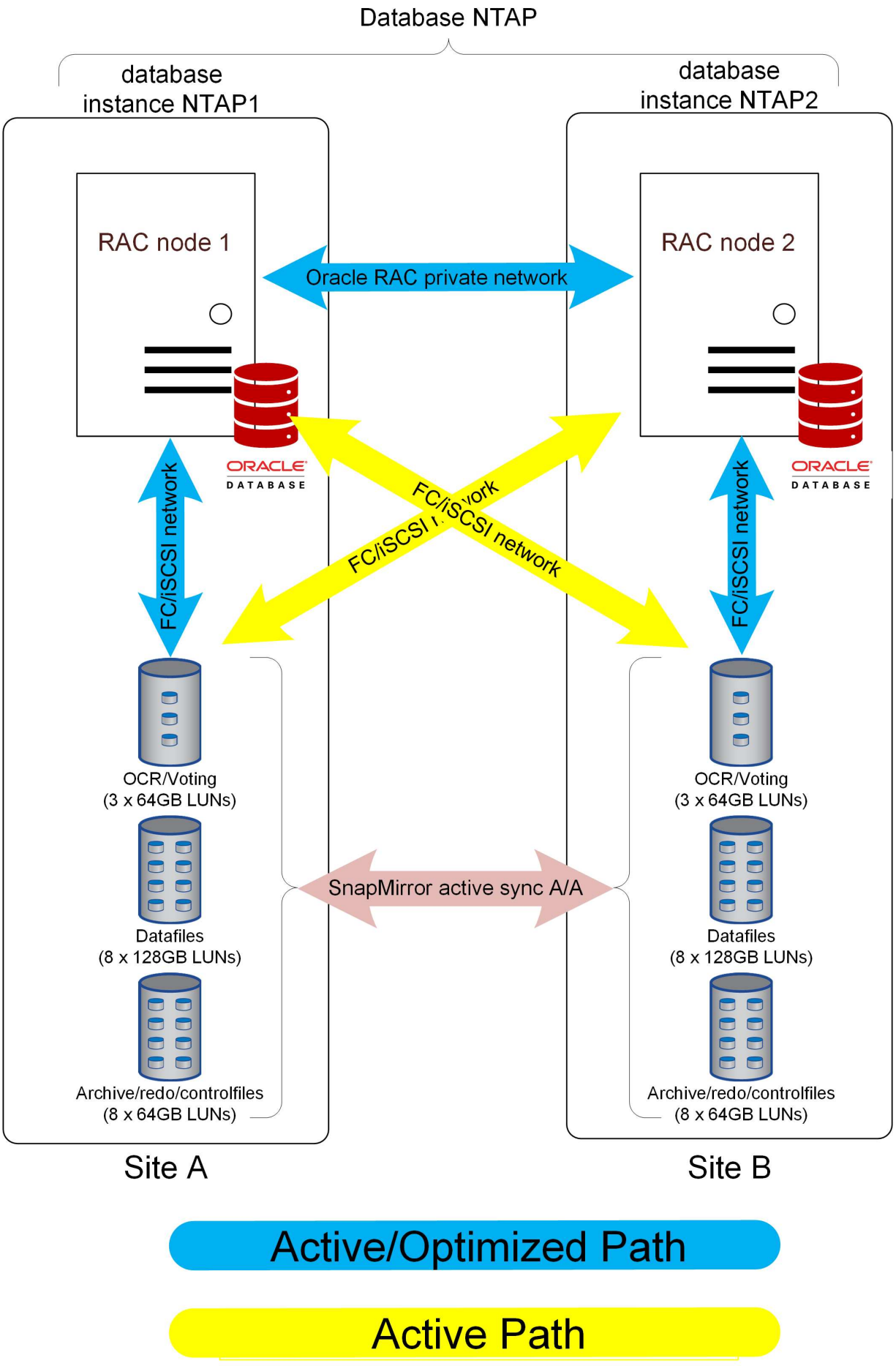
Para el acceso uniforme es necesario ampliar la SAN a través de varios sitios. La ventaja principal es que la pérdida de un sistema de almacenamiento no provocará la pérdida de una instancia de la base de datos. En su lugar, provocaría un cambio de multivía en el que se usan las rutas actualmente.

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

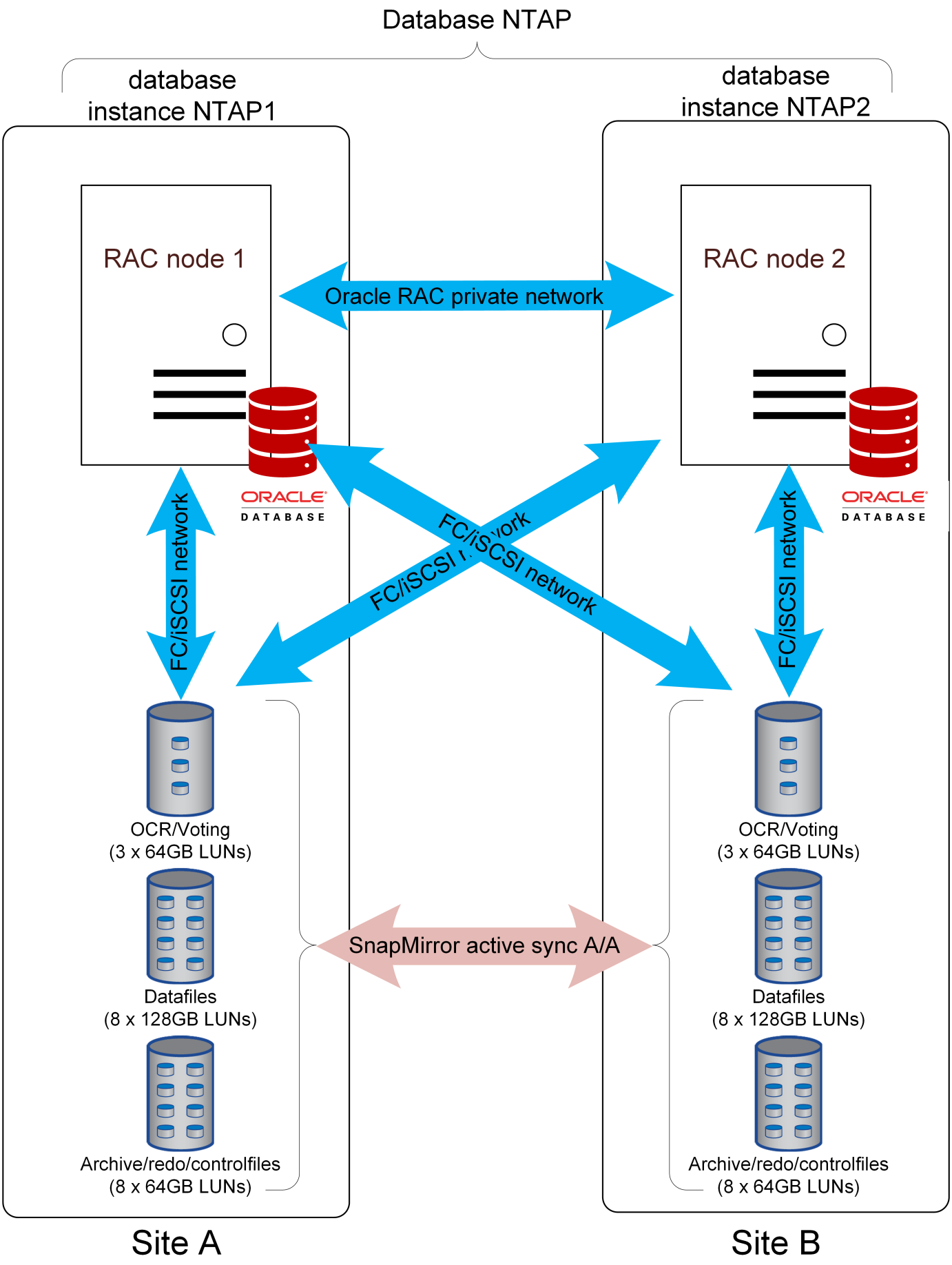
Si hay una latencia significativa entre los sitios, los sistemas AFF se pueden configurar con los ajustes de proximidad del host. Esto permite que cada sistema de almacenamiento tenga en cuenta qué hosts son locales y remotos y asigne prioridades de ruta según corresponda.


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

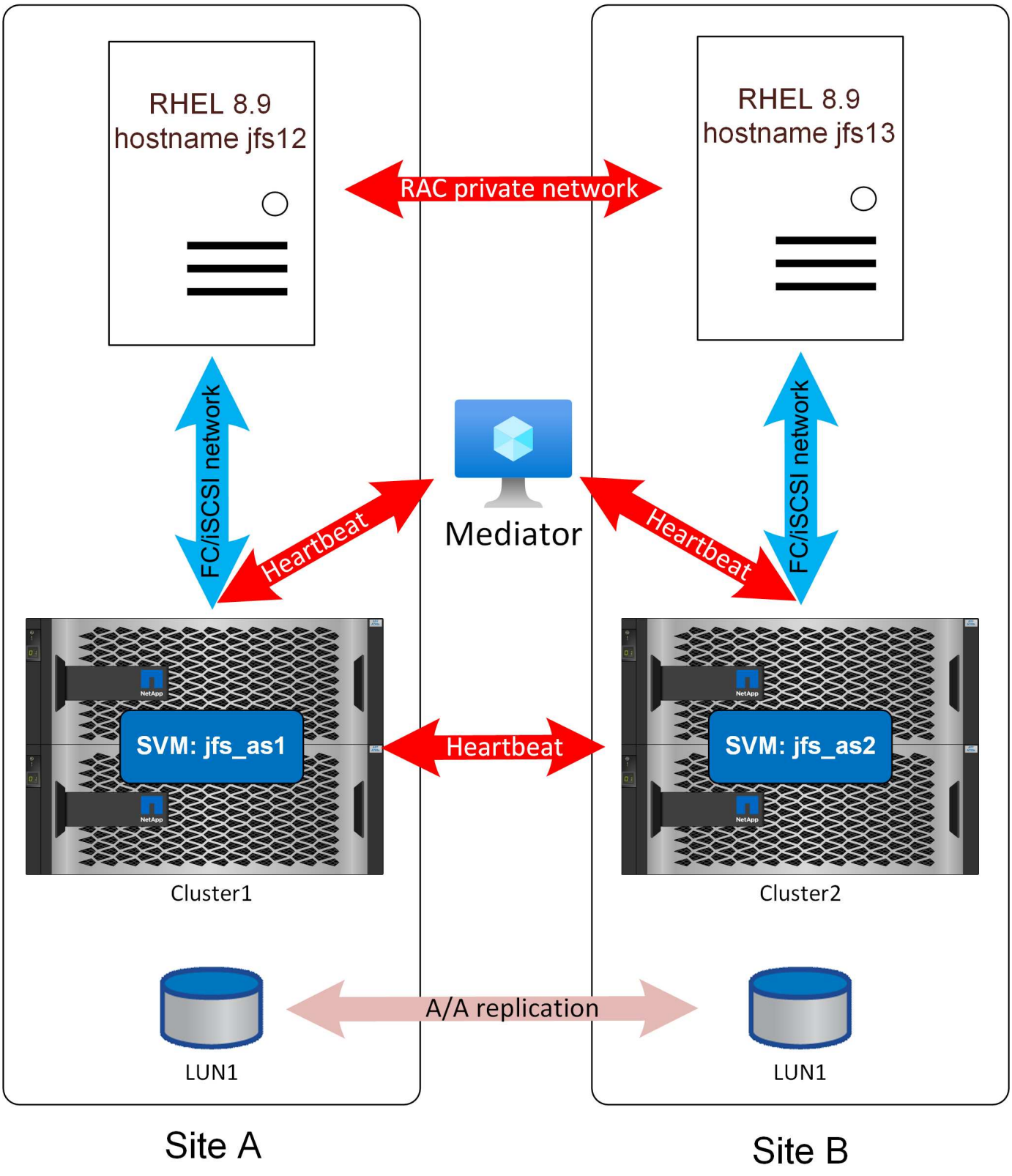
Cambiar el sitio preferido requiere una operación simple. I/O se detendrá durante un segundo o dos como autoridad sobre los cambios en el comportamiento de replicación entre los clústeres, pero I/O de otro modo no se verá afectado.

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.

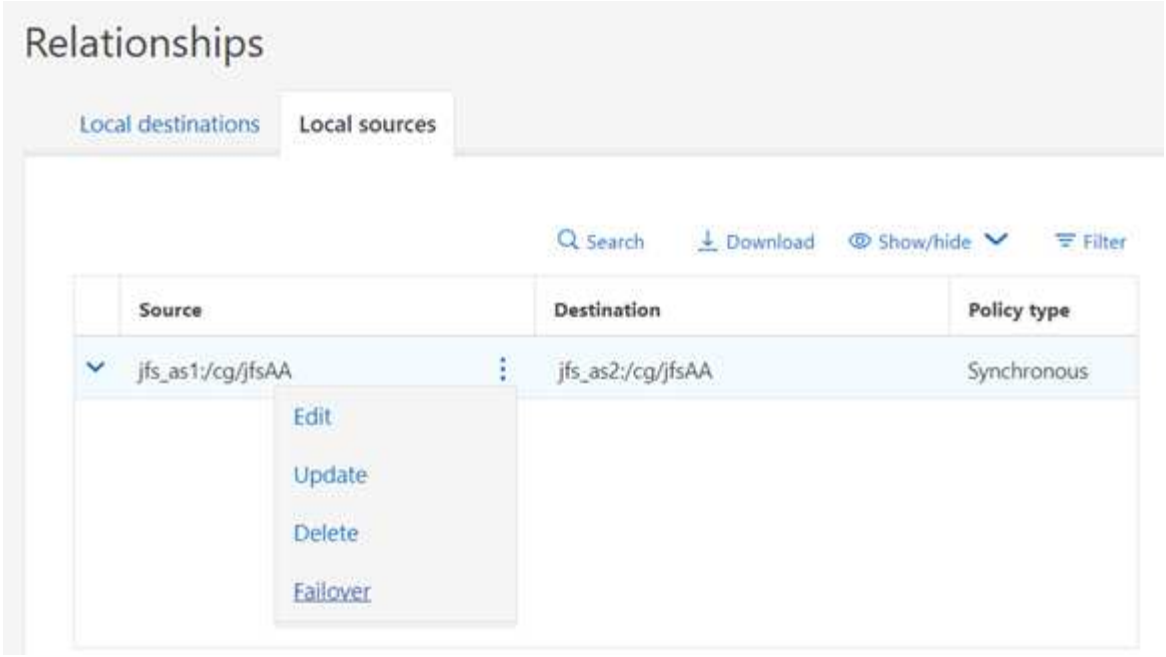
Recuperación manual tras fallos

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
```

Migración de bases de datos de Oracle

Descripción general

El aprovechamiento de las funciones de una nueva plataforma de almacenamiento tiene un requisito inevitable: Los datos deben estar situados en el nuevo sistema de almacenamiento. ONTAP simplifica el proceso de migración, lo que incluye migraciones y actualizaciones de ONTAP a ONTAP, importaciones de LUN externas y procedimientos para utilizar directamente el sistema operativo del host o el software de base de datos de Oracle.

Esta documentación sustituye al informe técnico *TR-4534: Migración de bases de datos de Oracle a sistemas de almacenamiento de NetApp*

En el caso de un nuevo proyecto de base de datos, no se trata de un problema, ya que los entornos de bases de datos y aplicaciones están construidos in situ. Sin embargo, la migración plantea desafíos especiales con

respecto a la interrupción del negocio, el tiempo necesario para completar la migración, las habilidades necesarias y la minimización de riesgos.

Scripts

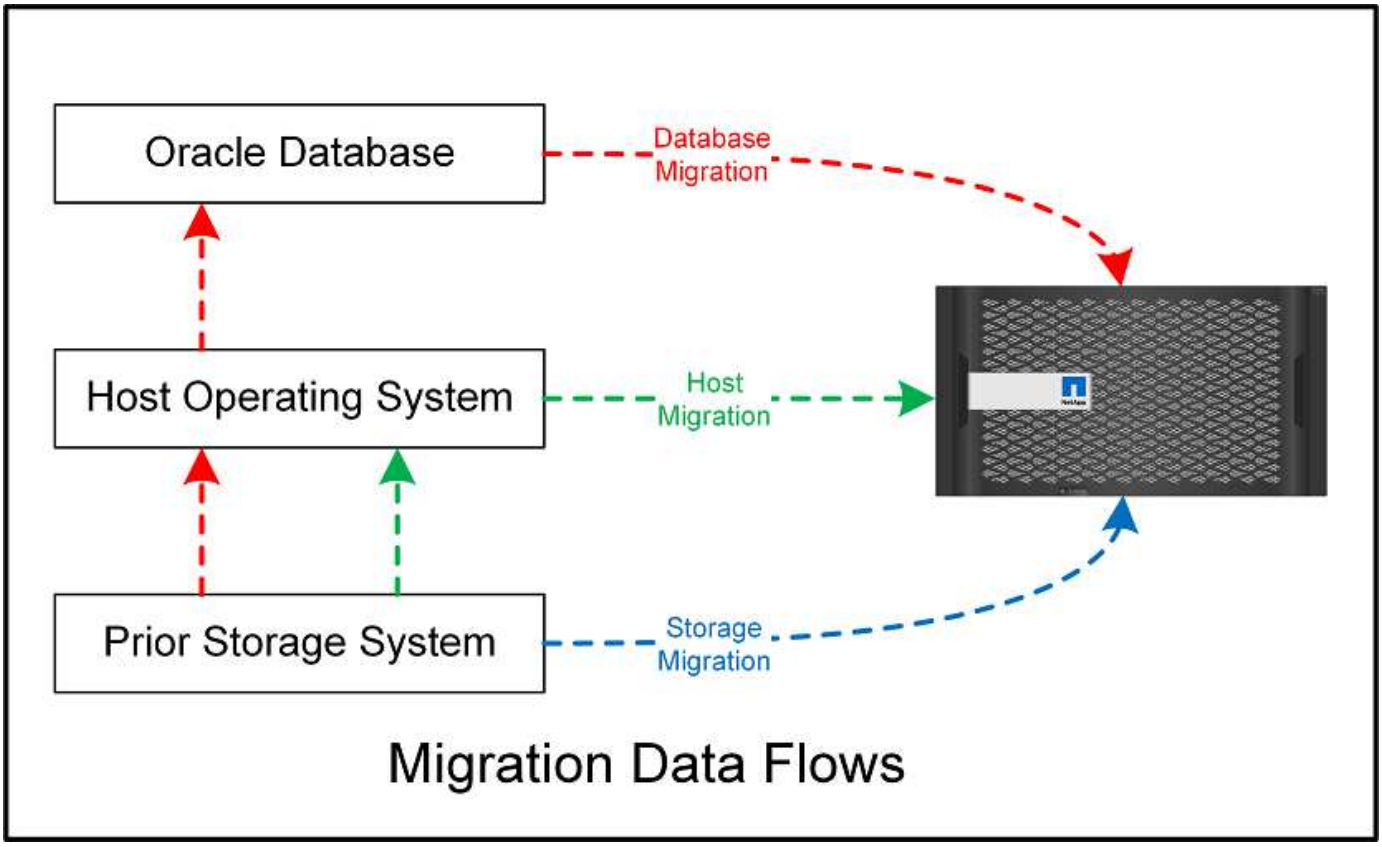
En esta documentación se proporcionan secuencias de comandos de ejemplo. Estos scripts proporcionan métodos de ejemplo de automatización de diversos aspectos de la migración para reducir la posibilidad de errores de usuario. Las secuencias de comandos pueden reducir las demandas generales sobre el PERSONAL DE TI responsable de la migración y pueden acelerar el proceso general. Todos estos scripts se extraen de los proyectos de migración reales realizados por los servicios profesionales de NetApp y los partners de NetApp. A lo largo de esta documentación se muestran ejemplos de su uso.

Planificación de migración

La migración de datos de Oracle puede producirse en uno de tres niveles: La base de datos, el host o la cabina de almacenamiento.

Las diferencias residen en qué componente de la solución general es responsable del movimiento de datos: La base de datos, el sistema operativo del host o el sistema de almacenamiento.

La siguiente figura muestra un ejemplo de los niveles de migración y el flujo de datos. En el caso de la migración a nivel de base de datos, los datos se mueven desde el sistema de almacenamiento original a través de las capas de base de datos y host al nuevo entorno. La migración al nivel de host es similar, pero los datos no pasan a través de la capa de aplicaciones y, en su lugar, se escriben en la nueva ubicación mediante procesos de host. Por último, con la migración a nivel del almacenamiento, una cabina como un sistema NetApp FAS es responsable del movimiento de datos.



Una migración a nivel de base de datos generalmente hace referencia al uso del envío de logs de Oracle a través de una base de datos en espera para completar una migración en la capa de Oracle. Las migraciones a

nivel de host se realizan utilizando la capacidad nativa de la configuración del sistema operativo host. Esta configuración incluye operaciones de copia de archivos mediante comandos como cp, tar y Oracle Recovery Manager (RMAN) o mediante un gestor de volúmenes lógicos (LVM) para reubicar los bytes subyacentes de un sistema de archivos. Oracle Automatic Storage Management (ASM) se clasifica como una capacidad de nivel de host porque se ejecuta por debajo del nivel de la aplicación de base de datos. ASM sustituye al administrador de volúmenes lógicos habitual en un host. Por último, los datos pueden migrarse a nivel de cabina de almacenamiento, lo cual significa que se encuentra debajo del nivel del sistema operativo.

Consideraciones DE PLANIFICACIÓN

La mejor opción para la migración depende de una combinación de factores, como la escala del entorno que se va a migrar, la necesidad de evitar el tiempo de inactividad y el esfuerzo general requerido para realizar la migración. Obviamente, las bases de datos grandes requieren más tiempo y esfuerzo para la migración, pero la complejidad de estas migraciones es mínima. Las bases de datos pequeñas se pueden migrar rápidamente, pero, si hay miles que migrar, la escala del esfuerzo puede crear complicaciones. Por último, cuanto mayor sea la base de datos, más probabilidades hay de que sea crítica para el negocio, lo cual da lugar a la necesidad de minimizar los tiempos de inactividad a la vez que se conserva una ruta de back-out.

Aquí se tratan algunas de las consideraciones para planificar una estrategia de migración.

Tamaño de datos

Los tamaños de las bases de datos que se migrarán afectan obviamente a la planificación de la migración, aunque el tamaño no afecta necesariamente al tiempo de transición. Cuando es necesario migrar una gran cantidad de datos, la principal cuestión es el ancho de banda. Las operaciones de copia suelen realizarse con I/O secuenciales eficientes Según estimaciones conservadoras, asuma el aprovechamiento del 50% del ancho de banda de red disponible para operaciones de copia. Por ejemplo, un puerto FC de 8GB Gb puede transferir aproximadamente 800Mbps Gb en teoría. Suponiendo una utilización del 50%, se puede copiar una base de datos a una velocidad de aproximadamente 400Mbps KB. Por lo tanto, una base de datos de 10TB TB se puede copiar en unas siete horas a esta velocidad.

La migración a distancias más largas generalmente requiere un enfoque más creativo, como el proceso de envío de registros explicado en "Movimiento de archivos de datos en línea". Las redes IP de larga distancia rara vez tienen ancho de banda en cualquier lugar cercano a las velocidades LAN o SAN. En un caso, NetApp ayudó en la migración a larga distancia de una base de datos de 220TB con tasas muy altas de generación de registros de archivo. El enfoque elegido para la transferencia de datos era el envío diario de cintas, ya que este método ofrecía el máximo ancho de banda posible.

Recuento de bases de datos

En muchos casos, el problema de mover una gran cantidad de datos no es el tamaño de los datos, sino la complejidad de la configuración que soporta la base de datos. No basta con saber que deben migrarse 50TB TB de bases de datos. Podría ser una única base de datos de misión crítica de 50TB TB, una colección de 4, 000 bases de datos heredadas o una combinación de datos de producción y no de producción. En algunos casos, gran parte de los datos se componen de clones de una base de datos de origen. Estos clones no tienen que migrarse de ninguna manera, ya que pueden volver a crearse fácilmente, especialmente cuando la nueva arquitectura está diseñada para aprovechar los volúmenes FlexClone de NetApp.

Para la planificación de la migración, hay que entender cuántas bases de datos están incluidas y cómo deben priorizarse. A medida que aumenta el número de bases de datos, la opción de migración preferida tiende a ser más baja y más baja en la pila. Por ejemplo, la copia de una única base de datos se puede realizar fácilmente con RMAN y una interrupción breve. Es la replicación a nivel de host.

Si hay bases de datos 50, es posible que sea más fácil evitar configurar una nueva estructura del sistema de archivos para recibir una copia de RMAN y, en su lugar, mover los datos. Este proceso puede realizarse

aprovechando la migración de LVM basada en host para reubicar datos de las LUN antiguas a nuevas LUN. De este modo, se traslada la responsabilidad del equipo del administrador de la base de datos (DBA) al equipo del sistema operativo y, como resultado, los datos se migran de forma transparente con respecto a la base de datos. La configuración del sistema de archivos no cambia.

Por último, si es necesario migrar 500 bases de datos en 200 servidores, pueden utilizarse opciones basadas en almacenamiento como la funcionalidad Importación de LUN externas (FLI) de ONTAP para realizar una migración directa de las LUN.

Vuelva a crear los requisitos de la arquitectura

Normalmente, el diseño de un archivo de base de datos debe modificarse para aprovechar las funciones de la nueva cabina de almacenamiento; sin embargo, esto no siempre es así. Por ejemplo, las funciones de las cabinas all-flash EF-Series se dirigen principalmente al rendimiento SAN y la fiabilidad de SAN. En la mayoría de los casos, las bases de datos pueden migrarse a una cabina EF-Series sin tener en cuenta ninguna necesidad de distribución de los datos. Los únicos requisitos son el alto nivel de IOPS, la baja latencia y la sólida fiabilidad. Aunque existen prácticas recomendadas en relación con factores como la configuración de RAID o los pools de discos dinámicos, los proyectos EF-Series rara vez requieren cambios significativos en la arquitectura general de almacenamiento para aprovechar estas funciones.

Por el contrario, la migración a ONTAP generalmente requiere tener en cuenta el diseño de la base de datos para asegurarse de que la configuración final aporta el máximo valor. Por sí mismo, ONTAP ofrece muchas funciones para un entorno de base de datos, incluso sin ningún esfuerzo de arquitectura específico. Y lo que es más importante, ofrece la capacidad de migrar sin interrupciones a un nuevo hardware cuando el hardware actual llega al final de su vida útil. En términos generales, una migración a ONTAP es la última migración que se debería realizar. Se actualiza el hardware subsiguiente in situ y los datos se migran a los nuevos medios de forma no disruptiva.

Con un poco de planificación, aún hay más beneficios disponibles. Las consideraciones más importantes rodean el uso de instantáneas. Las copias Snapshot son la base para realizar backups, restauraciones de datos y operaciones de clonado casi instantáneas. Como ejemplo del potencial de las copias Snapshot, el uso más grande conocido es con una única base de datos de 996TB TB que se ejecuta en unas 250 LUN en 6 controladoras. Puede realizarse backup de esta base de datos en 2 minutos, restaurarse en 2 minutos y clonarse en 15 minutos. Entre otras ventajas, se incluyen la capacidad de mover datos por el clúster en respuesta a los cambios en la carga de trabajo y la aplicación de controles de calidad de servicio para proporcionar un buen rendimiento constante en un entorno multibase de datos.

Tecnologías como los controles de calidad de servicio, la reubicación de datos, las snapshots y el clonado funcionan en prácticamente cualquier configuración. Sin embargo, generalmente se requiere algo de pensamiento para maximizar los beneficios. En algunos casos, la distribución del almacenamiento de la base de datos puede requerir cambios en el diseño para maximizar la inversión en la nueva cabina de almacenamiento. Estos cambios de diseño pueden afectar a la estrategia de migración, ya que las migraciones basadas en host o basadas en almacenamiento replican la distribución de datos original. Podrían ser necesarios pasos adicionales para completar la migración y ofrecer una distribución de datos optimizada para ONTAP. Los procedimientos que se muestran en la "Descripción general de los procedimientos de migración de Oracle" y más tarde, mostrar algunos de los métodos no solo para migrar una base de datos, sino para migrarla al diseño final óptimo con el mínimo esfuerzo.

Tiempo de transición

Se debe determinar la interrupción máxima permitida del servicio durante la transición. Es un error común asumir que todo el proceso de migración provoca interrupciones. Muchas tareas pueden completarse antes de que comience cualquier interrupción del servicio y muchas opciones permiten completar la migración sin interrupciones ni interrupciones del servicio. Incluso cuando resulte imposible evitar las interrupciones, debe definir la interrupción del servicio máxima permitida, puesto que la duración del tiempo de transición varía de

un procedimiento a otro.

Por ejemplo, la copia de una base de datos de 10TB GB normalmente requiere aproximadamente siete horas para completarse. Si su empresa necesita permitir un fallo de siete horas, la copia de archivos es una opción fácil y segura para la migración. Si cinco horas son inaceptables, un simple proceso de envío de registros (consulte "Envío de registros de Oracle") puede configurarse con un esfuerzo mínimo para reducir el tiempo de transición a aproximadamente 15 minutos. Durante este tiempo, un administrador de la base de datos puede completar el proceso. Si 15 minutos son inaceptables, el proceso final de transición se puede automatizar mediante secuencias de comandos para reducir el tiempo de transición a tan solo unos minutos. Siempre se puede acelerar la migración, pero hacerlo conlleva un coste de tiempo y esfuerzo. Los objetivos de tiempo de transición deben basarse en lo que sea aceptable para la empresa.

Ruta de retroceso

Ninguna migración está completamente exenta de riesgos. Incluso si la tecnología funciona perfectamente, siempre existe la posibilidad de error del usuario. El riesgo asociado a una ruta de migración elegida debe tenerse en cuenta junto con las consecuencias de una migración fallida. Por ejemplo, la capacidad transparente de migración de almacenamiento en línea de Oracle ASM es una de sus funciones clave, y este método es una de las más fiables conocidas. Sin embargo, los datos se copian de forma irreversible con este método. En el caso muy poco probable de que se produzca un problema con ASM, no hay una ruta de salida fácil. La única opción es restaurar el entorno original o utilizar ASM para revertir la migración de nuevo a las LUN originales. El riesgo puede minimizarse, pero no eliminarse, realizando un backup del tipo snapshot en el sistema de almacenamiento original, asumiendo que el sistema sea capaz de realizar dicha operación.

Ensayo

Algunos procedimientos de migración deben verificarse por completo antes de la ejecución. La necesidad de migración y ensayo del proceso de transición es una solicitud común con bases de datos críticas para la misión para la que la migración debe tener éxito y se debe minimizar el tiempo de inactividad. Además, las pruebas de aceptación del usuario se incluyen con frecuencia como parte del trabajo posterior a la migración y el sistema en general solo puede volver a la producción una vez que se hayan completado estas pruebas.

Si hay una necesidad de ensayo, varias capacidades de ONTAP pueden hacer el proceso mucho más fácil. En particular, las copias Snapshot pueden restablecer un entorno de prueba y crear rápidamente varias copias con gestión eficiente del espacio de un entorno de base de datos.

Procedimientos

Descripción general

Hay muchos procedimientos disponibles para la base de datos de migración de Oracle. El correcto depende de las necesidades de su empresa.

En muchos casos, los administradores de sistemas y los administradores de bases de datos cuentan con sus propios métodos preferidos para reubicar datos de volúmenes físicos, realizar mirroring y desduplicación o utilizar Oracle RMAN para copiar datos.

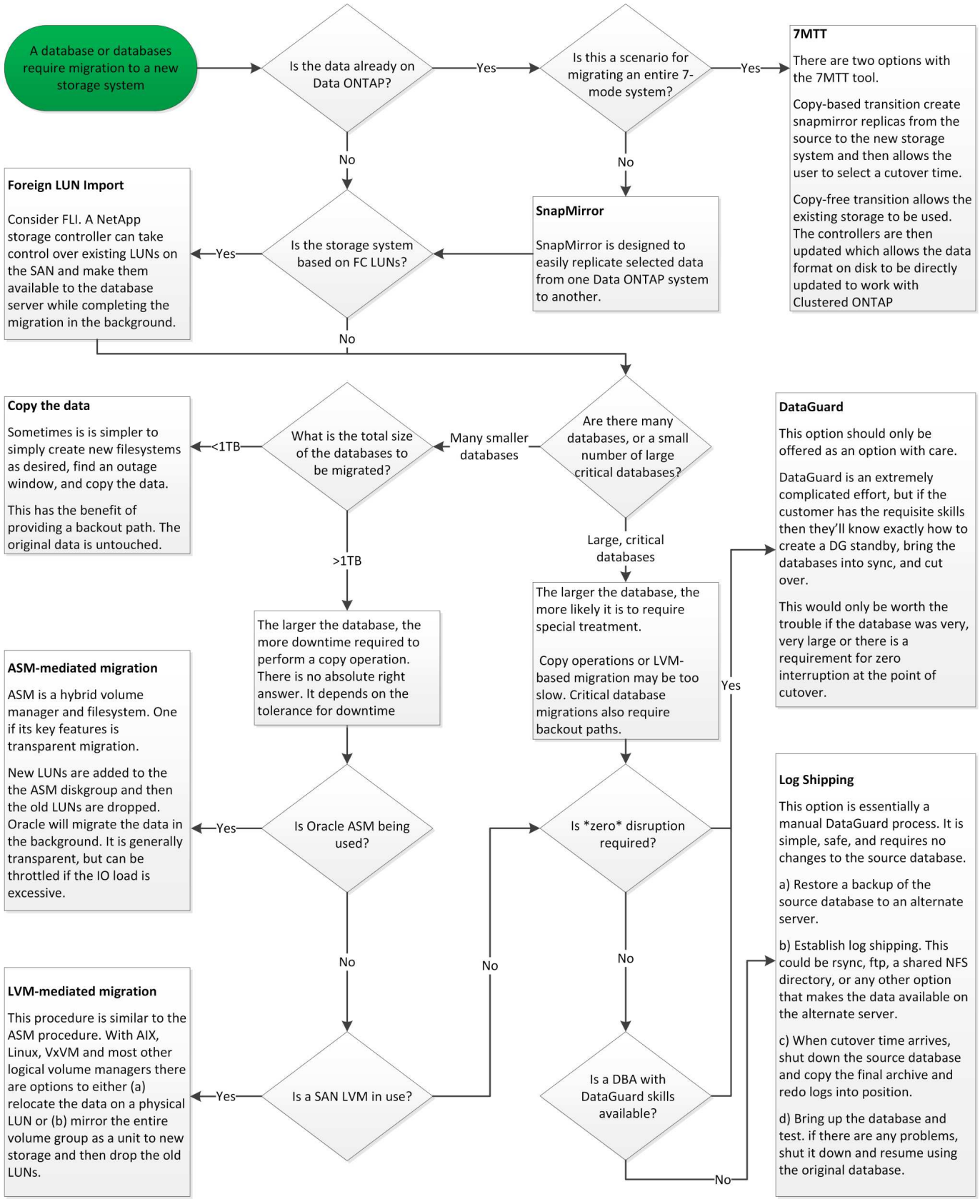
Estos procedimientos se proporcionan principalmente como orientación para el PERSONAL DE TI menos familiarizado con algunas de las opciones disponibles. Además, los procedimientos muestran las tareas, los requisitos de tiempo y las demandas de habilidades para cada método de migración. De este modo, otras partes como NetApp y los servicios profesionales de partners o el equipo de gestión de TI pueden apreciar de forma más completa los requisitos de cada procedimiento.

No existe una práctica recomendada única para crear una estrategia de migración. La creación de un plan

requiere primero comprender las opciones de disponibilidad y luego seleccionar el método que mejor se adapte a las necesidades del negocio. La siguiente figura ilustra las consideraciones básicas y las conclusiones típicas de los clientes, pero no es universalmente aplicable a todas las situaciones.

Por ejemplo, un paso plantea el problema del tamaño total de la base de datos. El siguiente paso depende de si la base de datos es mayor o menor que 1TB. Los pasos recomendados son simplemente eso: Recomendaciones basadas en las prácticas típicas del cliente. La mayoría de los clientes no utilizarían DataGuard para copiar una base de datos pequeña, pero algunos podrían. La mayoría de los clientes no intentarían copiar una base de datos de 50TB GB debido al tiempo necesario, pero algunos pueden tener una ventana de mantenimiento lo suficientemente grande como para permitir dicha operación.

El siguiente diagrama de flujo muestra los tipos de consideraciones sobre la ruta de migración que es mejor. Puede hacer clic con el botón derecho en la imagen y abrirla en una nueva pestaña para mejorar la legibilidad.



.

Movimiento de archivos de datos en línea

Oracle 12cR1 y las versiones superiores incluyen la capacidad de mover un archivo de datos mientras la base de datos permanece en línea. Además, funciona entre diferentes tipos de sistemas de archivos. Por ejemplo,

un archivo de datos se puede reubicar de un sistema de archivos xfs a ASM. Este método no se utiliza generalmente a escala debido al número de operaciones de movimiento de archivos de datos individuales que serían necesarias, pero es una opción que vale la pena considerar con bases de datos más pequeñas con menos archivos de datos.

Además, simplemente mover un archivo de datos es una buena opción para migrar partes de bases de datos existentes. Por ejemplo, los archivos de datos menos activos podrían reubicarse en un almacenamiento más rentable, como un volumen FabricPool que pueda almacenar bloques inactivos en el almacén de objetos.

Migración a nivel de base de datos

La migración a nivel de base de datos implica permitir que la base de datos vuelva a ubicar los datos. Específicamente, esto significa el envío de registros. Tecnologías como RMAN y ASM son productos de Oracle, pero, para la migración, funcionan en el nivel de host en el que copian archivos y gestionan volúmenes.

Trasvase de registros

La base para la migración a nivel de base de datos es el archive log de Oracle, que contiene un log de los cambios realizados en la base de datos. La mayoría de las veces, un registro de archivo forma parte de una estrategia de backup y recuperación. El proceso de recuperación comienza con la restauración de una base de datos y luego la reproducción de uno o más registros de archivos para que la base de datos alcance el estado deseado. Esta misma tecnología básica se puede usar para realizar una migración con poca o ninguna interrupción de las operaciones. Y lo que es más importante, esta tecnología permite la migración sin modificar la base de datos original, lo que mantiene un camino de back-out.

El proceso de migración comienza con la restauración de un backup de base de datos a un servidor secundario. Puede hacerlo de varias formas, pero la mayoría de los clientes utilizan su aplicación de backup normal para restaurar los archivos de datos. Después de restaurar los archivos de datos, los usuarios establecen un método para el envío de registros. El objetivo es crear una fuente constante de los archive logs generados por la base de datos primaria y reproducirlos en la base de datos restaurada para mantenerlos cerca del mismo estado. Cuando llega el tiempo de transposición, la base de datos de origen se cierra por completo y los archive logs finales, y en algunos casos los redo logs, se copian y se vuelven a reproducir. Es fundamental que los redo logs también se tengan en cuenta porque pueden contener algunas de las transacciones finales confirmadas.

Después de transferir y reproducir estos registros, ambas bases de datos son coherentes entre sí. En este momento, la mayoría de los clientes realizan algunas pruebas básicas. Si se produce algún error durante el proceso de migración, la reproducción de log debe informar de los errores y fallar. Aún es aconsejable realizar algunas pruebas rápidas basadas en consultas conocidas o actividades controladas por aplicaciones para verificar que la configuración es óptima. También es una práctica común crear una tabla de prueba final antes de cerrar la base de datos original para verificar si está presente en la base de datos migrada. Este paso garantiza que no se hayan producido errores durante la sincronización del registro final.

Una simple migración de envío de registros se puede configurar fuera de banda con respecto a la base de datos original, lo que la hace particularmente útil para las bases de datos de misión crítica. No se requieren cambios de configuración para la base de datos de origen, y la restauración y configuración inicial del entorno de migración no afectan a las operaciones de producción. Después de configurar el envío de registros, coloca algunas demandas de E/S en los servidores de producción. Sin embargo, el envío de registros consiste en lecturas secuenciales simples de los archive logs, lo que es poco probable que afecte al rendimiento de la base de datos de producción.

El envío de registros ha demostrado ser particularmente útil para proyectos de migración de larga distancia y alta tasa de cambio. En un ejemplo, una sola base de datos de 220TB TB se migró a una nueva ubicación aproximadamente a 500 kilómetros de distancia. La tasa de cambio fue extremadamente alta y las

restricciones de seguridad impidieron el uso de una conexión de red. El envío de registros se realizó mediante cinta y mensajería. Se restauró inicialmente una copia de la base de datos de origen mediante los procedimientos descritos a continuación. A continuación, los registros se enviaron semanalmente por mensajería hasta el momento de la transición, cuando se entregó el conjunto final de cintas y se aplicaron los registros a la base de datos de réplica.

Oracle DataGuard

En algunos casos, se garantiza un entorno DataGuard completo. No es correcto utilizar el término DataGuard para hacer referencia a cualquier envío de log o configuración de base de datos en espera. Oracle DataGuard es un marco completo para gestionar la replicación de bases de datos, pero no es una tecnología de replicación. La principal ventaja de un entorno DataGuard completo en un esfuerzo de migración es el switchover transparente de una base de datos a otra. DATAGUARD también permite un switchover transparente a la base de datos original si se detecta un problema, como un problema de rendimiento o conectividad de red con el nuevo entorno. Un entorno DataGuard completamente configurado requiere la configuración no sólo de la capa de base de datos, sino también de las aplicaciones, de modo que las aplicaciones puedan detectar un cambio en la ubicación de la base de datos primaria. En general, no es necesario utilizar DataGuard para completar una migración, pero algunos clientes tienen una amplia experiencia en DataGuard interna y ya dependen de ella para el trabajo de migración.

Vuelva a diseñar la arquitectura

Como hemos visto anteriormente, aprovechar las funciones avanzadas de las cabinas de almacenamiento en ocasiones requiere cambiar el diseño de la base de datos. Además, un cambio en el protocolo de almacenamiento, como migrar de ASM a un sistema de archivos NFS, altera necesariamente la distribución del sistema de archivos.

Una de las principales ventajas de los métodos de envío de registros, incluido DataGuard, es que el destino de replicación no tiene que coincidir con el origen. No hay problemas con el uso de un enfoque de envío de logs para migrar de ASM a un sistema de archivos normal o viceversa. El diseño preciso de los archivos de datos se puede cambiar en el destino para optimizar el uso de la tecnología de base de datos conectable (PDB) o para establecer controles de QoS de forma selectiva en ciertos archivos. En otras palabras, un proceso de migración basado en el envío de registros le permite optimizar el diseño de almacenamiento de la base de datos de forma fácil y segura.

Recursos del servidor

La necesidad de un segundo servidor es una limitación para la migración a nivel de base de datos. Hay dos maneras de usar este segundo servidor:

1. Puede utilizar el segundo servidor como nuevo directorio raíz permanente para la base de datos.
2. Puede utilizar el segundo servidor como servidor temporal. Una vez completada y probada la migración de datos a la nueva cabina de almacenamiento, los sistemas de archivos LUN o NFS se desconectan del servidor provisional y se vuelven a conectar al servidor original.

La primera opción es la más fácil, pero su uso podría no ser factible en entornos muy grandes que requieran servidores muy potentes. La segunda opción requiere trabajo adicional para volver a ubicar los sistemas de archivos en la ubicación original. Esta operación puede ser sencilla en la que NFS se utiliza como protocolo de almacenamiento, ya que los sistemas de archivos se pueden desmontar del servidor de almacenamiento provisional y volver a montarse en el servidor original.

Los sistemas de archivos basados en bloques requieren trabajo adicional para actualizar la división en zonas de FC o los iniciadores de iSCSI. Con la mayoría de los administradores de volúmenes lógicos (incluido ASM), los LUN se detectan automáticamente y se conectan después de que estén disponibles en el servidor original.

Sin embargo, algunas implementaciones de sistemas de archivos y LVM pueden requerir más trabajo para exportar e importar los datos. El procedimiento preciso puede variar, pero generalmente es fácil establecer un procedimiento simple y repetible para completar la migración y volver a alojar los datos en el servidor original.

Aunque es posible configurar el envío de logs y replicar una base de datos en un entorno de servidor único, la nueva instancia debe tener un SID de proceso diferente para reproducir los logs. Es posible traer temporalmente la base de datos bajo un juego diferente de IDs de proceso con un SID diferente y cambiarla más tarde. Sin embargo, esta operación puede resultar en una gran cantidad de actividades de gestión complicadas y pone en riesgo al entorno de bases de datos de que se produzcan errores por parte del usuario.

Migración de nivel de host

Migrar datos a nivel de host significa utilizar el sistema operativo del host y las utilidades asociadas para completar la migración. Este proceso incluye cualquier utilidad que copie datos, incluidos Oracle RMAN y Oracle ASM.

Copiado de datos

No se debe subestimar el valor de una operación de copia simple. Las infraestructuras de red modernas pueden transferir datos a velocidades medidas en gigabytes por segundo y las operaciones de copia de archivos se basan en una eficiente E/S de lectura y escritura secuencial Una operación de copia de host no puede evitar más interrupciones cuando se compara con el envío de registros, pero una migración supone algo más que movimiento de datos. Por lo general, incluye cambios en las redes, el tiempo de reinicio de la base de datos y las pruebas posteriores a la migración.

El tiempo real necesario para copiar los datos puede no ser significativo. Además, una operación de copia conserva una ruta de back-out garantizada, ya que los datos originales permanecen sin tocar. Si se produce algún problema durante el proceso de migración, se pueden volver a activar los sistemas de archivos originales con los datos originales.

Cambio de la plataforma

El cambio de plataforma hace referencia a un cambio en el tipo de CPU. Cuando una base de datos se migra desde una plataforma tradicional Solaris, AIX o HP-UX a x86 Linux, los datos se deben volver a formatear debido a los cambios en la arquitectura de la CPU. Las CPU SPARC, IA64 y POWER se conocen como procesadores big endian, mientras que las arquitecturas x86 y x86_64 se conocen como little endian. Como resultado, algunos datos de los archivos de datos de Oracle se ordenan de forma diferente dependiendo del procesador en uso.

Tradicionalmente, los clientes utilizaban DataPump para replicar datos entre plataformas. DataPump es una utilidad que crea un tipo especial de exportación de datos lógicos que se puede importar más rápidamente en la base de datos destino. Debido a que crea una copia lógica de los datos, DataPump deja atrás las dependencias de endianness del procesador. Algunos clientes siguen utilizando DataPump para la transformación de plataformas, pero se ha puesto a disposición una opción más rápida con los tablespaces transportables multiplataforma de Oracle 11g:. Este avance permite que un tablespace se convierta a un formato endian diferente. Se trata de una transformación física que ofrece un mejor rendimiento que una exportación de DataPump, que debe convertir bytes físicos en datos lógicos y luego volver a convertir a bytes físicos.

No se trata completamente de la NetApp documentación de DataPump y los espacios de tablas transportables. No obstante, NetApp cuenta con algunas recomendaciones basadas en nuestra experiencia al ayudar a los clientes durante la migración a un nuevo registro de cabina de almacenamiento con una nueva arquitectura de CPU:

- Si se utiliza DataPump, el tiempo necesario para completar la migración se debe medir en un entorno de prueba. A veces, los clientes se sorprenden por el momento necesario para completar la migración. Este tiempo de inactividad adicional inesperado puede provocar una interrupción.
- Muchos clientes creen erróneamente que los tablespaces transportables entre plataformas no requieren conversión de datos. Cuando se utiliza una CPU con un endian diferente, un RMAN `convert` la operación debe realizarse en los archivos de datos de antemano. No se trata de una operación instantánea. En algunos casos, el proceso de conversión se puede acelerar al tener varios subprocesos que funcionan en diferentes archivos de datos, pero el proceso de conversión no se puede evitar.

Migración controlada por el gestor de volúmenes lógicos

Los LVM funcionan tomando un grupo de uno o más LUN y dividiéndolos en unidades pequeñas que normalmente se conocen como extensiones. El pool de extensiones se utiliza entonces como origen para crear volúmenes lógicos que están esencialmente virtualizados. Esta capa de virtualización proporciona valor de varias formas:

- Los volúmenes lógicos pueden utilizar extensiones extraídas de varios LUN. Cuando se crea un sistema de archivos en un volumen lógico, puede utilizar todas las funcionalidades de rendimiento de todas las LUN. También promueve la carga uniforme de todas las LUN en el grupo de volúmenes, lo que ofrece un rendimiento más previsible.
- Los volúmenes lógicos se pueden cambiar de tamaño agregando y, en algunos casos, eliminando extensiones. Cambiar el tamaño de un sistema de archivos en un volumen lógico suele ser no disruptivo.
- Los volúmenes lógicos pueden migrarse de forma no disruptiva moviendo las extensiones subyacentes.

La migración mediante un LVM funciona de dos maneras: Mover una extensión o duplicar/desactivar una extensión. La migración de LVM utiliza I/O secuencial de grandes bloques y solo rara vez crea preocupación sobre el rendimiento. Si esto se convierte en un problema, normalmente existen opciones para reducir la tasa de I/O. Hacerlo, aumenta el tiempo necesario para completar la migración pero reduce la carga de I/O en el host y los sistemas de almacenamiento.

Retrovisor y retrovisor

Algunos administradores de volúmenes, como AIX LVM, permiten al usuario especificar el número de copias para cada extensión y controlar qué dispositivos alojan cada copia. La migración se lleva a cabo tomando un volumen lógico existente, reflejando las extensiones subyacentes a los nuevos volúmenes, esperando a que se sincronicen las copias y borrando la antigua. Si se desea una ruta de retroceso, se puede crear una instantánea de los datos originales antes del punto en el que se descarta la copia de duplicación. También puede apagar el servidor brevemente para enmascarar las LUN originales antes de eliminar forzosamente las copias de duplicación contenidas. De este modo se conserva una copia recuperable de los datos en su ubicación original.

Migración de extensiones

Casi todos los gestores de volúmenes permiten migrar extensiones y, a veces, existen varias opciones. Por ejemplo, algunos administradores de volúmenes permiten que un administrador reubique las extensiones individuales de un volumen lógico específico, de almacenamiento antiguo a nuevo. Los gestores de volúmenes, como Linux LVM2, ofrecen el `pvmove` Comando, que reubica todas las extensiones del dispositivo LUN especificado en una LUN nueva. Después de evacuar la LUN antigua, puede quitarse.

El principal riesgo para las operaciones es la eliminación de LUN antiguas y no utilizadas de la configuración. Debe tenerse mucho cuidado al cambiar la división en zonas de FC y eliminar los dispositivos LUN obsoletos.

Gestión Automática de Almacenamiento de Oracle

Oracle ASM es un gestor de volúmenes lógicos y un sistema de archivos combinados. En un nivel superior, Oracle ASM toma una colección de LUN, los divide en pequeñas unidades de asignación y los presenta como un único volumen conocido como grupo de discos ASM. ASM también incluye la capacidad de reflejar el grupo de discos mediante la definición del nivel de redundancia. Un volumen puede estar no reflejado (redundancia externa), reflejado (redundancia normal) o reflejado en tres direcciones (alta redundancia). Se debe tener cuidado al configurar el nivel de redundancia porque no se puede cambiar después de la creación.

ASM también proporciona la funcionalidad del sistema de archivos. Aunque el sistema de archivos no está visible directamente desde el host, la base de datos Oracle puede crear, mover y suprimir archivos y directorios en un grupo de discos ASM. Además, la estructura puede ser navegada usando la utilidad asmcmd.

Al igual que con otras implementaciones de LVM, Oracle ASM optimiza el rendimiento de E/S mediante la segmentación y el equilibrio de carga de E/S de cada archivo en todas las LUN disponibles. En segundo lugar, las extensiones subyacentes se pueden reubicar para permitir tanto el cambio de tamaño del grupo de discos de ASM como la migración. Oracle ASM automatiza el proceso mediante la operación de reequilibrio. Se agregan nuevos LUN a un grupo de discos ASM y se eliminan LUN antiguas, lo que activa la reubicación de extensiones y la posterior caída de la LUN evacuada del grupo de discos. Este proceso es uno de los métodos de migración más probados, y la fiabilidad de ASM a la hora de proporcionar una migración transparente es posiblemente su característica más importante.

Como el nivel de mirroring de Oracle ASM es fijo, no se puede utilizar con el método de migración mirror y demirror.

Migración de nivel de almacenamiento

La migración al nivel de almacenamiento significa realizar la migración por debajo tanto del nivel de aplicación como del sistema operativo. Anteriormente, esto suponía el uso de dispositivos especializados que copiaban LUN a nivel de red, pero estas funcionalidades ahora se encuentran de forma nativa en ONTAP.

SnapMirror

La migración de bases de datos desde sistemas NetApp se realiza casi universalmente con el software de replicación de datos SnapMirror de NetApp. El proceso implica configurar una relación de mirroring para los volúmenes que se migrarán, lo que permite que se sincronicen y luego esperar la ventana de transposición. Cuando llega, la base de datos de origen se cierra, se realiza una actualización de duplicación final y se interrumpe la duplicación. A continuación, los volúmenes de réplica están listos para su uso, ya sea montando un directorio de sistema de archivos NFS contenido o detectando los LUN contenidos e iniciando la base de datos.

La reubicación de volúmenes dentro de un único clúster de ONTAP no se considera una migración, sino una rutina `volume move` funcionamiento. SnapMirror se utiliza como motor de replicación de datos en el clúster. Este proceso está totalmente automatizado. No hay otros pasos de migración que se deben realizar cuando atributos del volumen, como la asignación de LUN o los permisos de exportación de NFS, se mueven con el propio volumen. La reubicación no provoca interrupciones en las operaciones del host. En algunos casos, el acceso a la red debe actualizarse para garantizar que se accede a los datos recién reubicados de la forma más eficiente posible, pero estas tareas también no producen interrupciones.

Importación de LUN externa (FLI)

FLI es una función que permite que un sistema Data ONTAP que ejecuta 8,3 o superior migre un LUN existente desde otra cabina de almacenamiento. El procedimiento es simple: El sistema ONTAP se divide en

zonas en la cabina de almacenamiento existente como si fuera cualquier otro host SAN. A continuación, Data ONTAP toma el control de las LUN heredadas deseadas y migra los datos subyacentes. Además, el proceso de importación utiliza la configuración de eficiencia del volumen nuevo a medida que se migran los datos, lo que significa que los datos se pueden comprimir y deduplicar online durante el proceso de migración.

La primera implementación de FLI en Data ONTAP 8,3 solo permitía la migración sin conexión. Esta transferencia fue extremadamente rápida, pero seguía significando que los datos de la LUN no estaban disponibles hasta que se completó la migración. La migración en línea se introdujo en Data ONTAP 8,3.1. Este tipo de migración minimiza las interrupciones al permitir que ONTAP sirva datos de LUN durante el proceso de transferencia. Se produce una breve interrupción mientras se vuelve a dividir en zonas el host para usar los LUN a través de ONTAP. No obstante, tan pronto como se realicen estos cambios, los datos volverán a estar accesibles y seguirán siendo accesibles durante todo el proceso de migración.

La I/O de lectura se proxy mediante ONTAP hasta que se completa la operación de copia, mientras que la I/O de escritura se escribe de forma síncrona en el LUN externo y en el LUN de ONTAP. Las dos copias LUN se mantienen sincronizadas de esta manera hasta que el administrador ejecuta una transposición completa que libera la LUN externa y ya no replica las escrituras.

FLI está diseñado para funcionar con FC, pero si se desea cambiar a iSCSI, el LUN migrado puede volver a asignarse fácilmente como LUN iSCSI una vez finalizada la migración.

Entre las características de FLI se encuentra la detección y ajuste automático de alineación. En este contexto, el término alineación hace referencia a una partición en un dispositivo LUN. Para un rendimiento óptimo es necesario alinear las E/S con bloques de 4K KB. Si una partición se coloca en un desplazamiento que no es múltiplo de 4K, el rendimiento se ve afectado.

Hay un segundo aspecto de la alineación que no se puede corregir ajustando un desplazamiento de partición: El tamaño del bloque del sistema de archivos. Por ejemplo, un sistema de archivos ZFS generalmente toma por defecto un tamaño de bloque interno de 512 bytes. Otros clientes que usan AIX han creado ocasionalmente sistemas de archivos JFS2 con un tamaño de bloque de 512 o 1, 024 bytes. Aunque es posible que el sistema de archivos esté alineado con un límite de 4K KB, los archivos creados dentro de ese sistema de archivos no lo están y el rendimiento se resienta.

FLI no debe utilizarse en estas circunstancias. Aunque se puede acceder a los datos tras la migración, el resultado son sistemas de archivos con serias limitaciones de rendimiento. Como principio general, cualquier sistema de archivos que admita una carga de trabajo de sobrescritura aleatoria en ONTAP debería utilizar un tamaño de bloque de 4K KB. Esto es aplicable principalmente a cargas de trabajo como los archivos de datos de bases de datos e implementaciones de VDI. El tamaño de bloque se puede identificar mediante los comandos del sistema operativo del host relevantes.

Por ejemplo, en AIX, el tamaño de bloque se puede ver con `lsfs -q`. Con Linux, `xfs_info` y.. `tune2fs` se puede utilizar para `xfs` y.. `ext3/ext4`, respectivamente. Con `zfs`, el comando es `zdb -C`.

El parámetro que controla el tamaño del bloque es `ashift` y, por lo general, el valor predeterminado es 9, lo que significa 2^9, o 512 bytes. Para un rendimiento óptimo, el `ashift` El valor debe ser 12 (2^12=4K). Este valor se define en el momento en que se crea zpool y no se puede cambiar, lo que significa que los datos zpools con un `ashift` los datos que no sean 12 se deben migrar copiando a un zpool recién creado.

Oracle ASM no tiene un tamaño de bloque fundamental. El único requisito es que la partición en la que se crea el disco de ASM esté alineada correctamente.

Herramienta de transición de 7-Mode

La herramienta de transición de 7-Mode (7MTT) es una utilidad de automatización que se usa para migrar configuraciones de 7- Mode de gran tamaño a ONTAP. La mayoría de los clientes de bases de datos

encuentran otros métodos más sencillos, en parte, debido a que suelen migrar la base de datos de sus entornos por base de datos en lugar de reubicar todo el espacio físico de almacenamiento. Además, normalmente las bases de datos solo forman parte de un entorno de almacenamiento de mayor tamaño. Por tanto, las bases de datos suelen migrarse de forma individual y entonces el entorno restante puede moverse con el 7MTT.

Hay un número pequeño pero significativo de clientes que disponen de sistemas de almacenamiento dedicados a entornos de bases de datos complicados. Estos entornos pueden contener numerosos volúmenes, copias Snapshot y numerosos detalles de configuración, como permisos de exportación, grupos de iniciadores de LUN, permisos de usuario y configuración de protocolo ligero de acceso a directorios. En tales casos, las capacidades de automatización de 7MTT pueden simplificar una migración.

7MTT puede funcionar en uno de dos modos:

- **Transición basada en copia (CBT).** 7MTT Con CBT se configuran los volúmenes de SnapMirror a partir de un sistema 7-Mode existente en el nuevo entorno. Una vez que los datos están sincronizados, 7MTT orquesta el proceso de transición.
- **Transición sin copia (CFT).** 7MTT con CFT se basa en la conversión in situ de las bandejas de discos 7-Mode existentes. No se copian datos y las bandejas de discos existentes pueden volver a utilizarse. La configuración existente de la protección de datos y la eficiencia del almacenamiento se conserva.

La principal diferencia entre estas dos opciones es que la transición sin copias es un método muy importante, en el que todas las bandejas de discos conectadas al par de alta disponibilidad 7- Mode original deben reubicarse en el nuevo entorno. No existe una opción para mover un subconjunto de bandejas. El enfoque basado en copia permite mover los volúmenes seleccionados. También hay potencialmente un periodo de transición más largo con una transición sin copias debido al vínculo necesario para volver a conectar las bandejas de discos y convertir los metadatos. Según la experiencia práctica, NetApp recomienda permitir 1 hora para reubicar y reconectar las bandejas de discos, y entre 15 minutos y 2 horas para la conversión de metadatos.

Migración de archivos de datos

Los archivos de datos de Oracle individuales se pueden mover con un solo comando.

Por ejemplo, el siguiente comando mueve el archivo de datos IOPST.dbf del sistema de archivos `/oradata2` al sistema de archivos `/oradata3`.

```
SQL> alter database move datafile  '/oradata2/NTAP/IOPS002.dbf' to
'/oradata3/NTAP/IOPS002.dbf';
Database altered.
```

Mover un archivo de datos con este método puede ser lento, pero normalmente no debería producir suficientes E/S que interfiera con las cargas de trabajo diarias de la base de datos. Por el contrario, la migración a través del reequilibrio de ASM puede ejecutarse mucho más rápido, pero a costa de ralentizar la base de datos general mientras se mueven los datos.

El tiempo necesario para mover archivos de datos se puede medir fácilmente creando un archivo de datos de prueba y moviéndolo después. El tiempo transcurrido para la operación se registra en los datos de v$session:

```
SQL> set linesize 300;
SQL> select elapsed_seconds||':'||message from v$session_longops;
ELAPSED_SECONDS||':'||MESSAGE
------------------------------------------------------------------------
---------------
351:Online data file move: data file 8: 22548578304 out of 22548578304
bytes done
SQL> select bytes / 1024 / 1024 /1024 as GB from dba_data_files where
FILE_ID = 8;
        GB
----------
        21
```

En este ejemplo, el archivo que se movió era el archivo de datos 8, que tenía un tamaño de 21GB GB y requería unos 6 minutos para migrar. El tiempo necesario depende obviamente de las funcionalidades del sistema de almacenamiento, la red de almacenamiento y la actividad general de las bases de datos que se produzca en el momento de la migración.

Trasvase de registros

El objetivo de una migración mediante el envío de registros es crear una copia de los archivos de datos originales en una nueva ubicación y, a continuación, establecer un método de envío de cambios en el nuevo entorno.

Una vez establecido, el envío y la reproducción de registros se pueden automatizar para mantener la base de datos de réplicas en gran medida sincronizada con la fuente. Por ejemplo, se puede programar un trabajo cron para (a) copiar los logs más recientes en la nueva ubicación y (b) reproducirlos cada 15 minutos. De este modo, se genera una interrupción mínima en el momento de la transición, ya que no se deben volver a reproducir más de 15 minutos de registros de archivo.

El procedimiento que se muestra a continuación también es esencialmente una operación de clonado de base de datos. La lógica mostrada es similar al motor de NetApp SnapManager para Oracle (SMO) y el plugin para Oracle de NetApp SnapCenter. Algunos clientes han utilizado el procedimiento mostrado en los flujos de trabajo de WFA o en los scripts para operaciones de clonado personalizadas. Aunque este procedimiento es más manual que usar SMO o SnapCenter, todavía dispone de secuencias de comandos sencillas y las API de gestión de datos en ONTAP simplifican aún más el proceso.

Envío de registros: Sistema de archivos al sistema de archivos

Este ejemplo muestra la migración de una base de datos denominada WAFFLE de un sistema de archivos ordinario a otro sistema de archivos ordinario ubicado en un servidor diferente. También ilustra el uso de SnapMirror para realizar una copia rápida de los archivos de datos, pero esto no forma parte integral del procedimiento general.

Crear copia de seguridad de base de datos

El primer paso es crear una copia de seguridad de la base de datos. En concreto, este procedimiento requiere un juego de archivos de datos que se pueda utilizar para la reproducción del archive log.

Entorno Oracle

En este ejemplo, la base de datos de origen se encuentra en un sistema ONTAP. El método más sencillo para crear un backup de una base de datos es mediante una instantánea. La base de datos se coloca en modo de backup dinámico durante unos segundos mientras a. `snapshot create` la operación se ejecuta en el volumen que aloja los archivos de datos.

```
SQL> alter database begin backup;
Database altered.
```

```
Cluster01::*> snapshot create -vserver vserver1 -volume jfsc1_oradata
hotbackup
Cluster01::*>
```

```
SQL> alter database end backup;
Database altered.
```

El resultado es una instantánea en disco llamada `hotbackup` que contiene una imagen de los archivos de datos mientras se encuentra en modo de copia de seguridad activa. Si se combinan con los archive logs adecuados para que los archivos de datos sean coherentes, se pueden utilizar los datos de esta copia Snapshot como base de la restauración o el clon. En este caso, se replica en el nuevo servidor.

Restauración al nuevo entorno

La copia de seguridad se debe restaurar ahora en el nuevo entorno. Esto puede realizarse de varias maneras, incluida Oracle RMAN, restauración desde una aplicación de backup como NetBackup o una operación de copia sencilla de archivos de datos ubicados en modo de backup dinámico.

En este ejemplo, se usa SnapMirror para replicar el backup en caliente de la copia Snapshot en una nueva ubicación.

1. Cree un volumen nuevo para recibir los datos de las snapshots. Inicialice el mirroring a partir de `jfsc1_oradata` para `vol_oradata`.

   ```
   Cluster01::*> volume create -vserver vserver1 -volume vol_oradata
   -aggregate data_01 -size 20g -state online -type DP -snapshot-policy
   none -policy jfsc3
   [Job 833] Job succeeded: Successful
   ```

```
Cluster01::*> snapmirror initialize -source-path vserver1:jfsc1_oradata
-destination-path vserver1:vol_oradata
Operation is queued: snapmirror initialize of destination
"vserver1:vol_oradata".
Cluster01::*> volume mount -vserver vserver1 -volume vol_oradata
-junction-path /vol_oradata
Cluster01::*>
```

2. Una vez definido el estado mediante SnapMirror, que indica que la sincronización está completada, actualice el mirror según la snapshot que desee.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields state
source-path             destination-path        state
----------------------- ----------------------- ------------
vserver1:jfsc1_oradata vserver1:vol_oradata SnapMirrored
```

```
Cluster01::*> snapmirror update -destination-path vserver1:vol_oradata
-source-snapshot hotbackup
Operation is queued: snapmirror update of destination
"vserver1:vol_oradata".
```

3. La sincronización correcta se puede verificar en el `newest-snapshot` en el volumen de reflejo.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields newest-snapshot
source-path             destination-path        newest-snapshot
----------------------- ----------------------- ---------------
vserver1:jfsc1_oradata vserver1:vol_oradata hotbackup
```

4. El espejo puede romperse.

```
Cluster01::> snapmirror break -destination-path vserver1:vol_oradata
Operation succeeded: snapmirror break for destination
"vserver1:vol_oradata".
Cluster01::>
```

5. Monte el nuevo sistema de archivos.Con los sistemas de archivos basados en bloques, los procedimientos precisos varían según el LVM en uso. Debe configurarse la división en zonas de FC o las conexiones iSCSI. Después de establecer la conectividad a las LUN, comandos como Linux `pvscan` Puede que sea necesario detectar qué grupos de volúmenes o LUN tienen que estar correctamente configurados para que ASM pueda detectar.

En este ejemplo, se utiliza un sistema de archivos NFS simple. Este sistema de archivos se puede montar directamente.

```
fas8060-nfs1:/vol_oradata        19922944   1639360   18283584   9%
/oradata
fas8060-nfs1:/vol_logs            9961472       128    9961344   1%
/logs
```

Crear plantilla de creación de archivo de control

A continuación, debe crear una plantilla de archivo de control. La `backup controlfile to trace` command crea comandos de texto para volver a crear un archivo de control. Esta función puede ser útil para restaurar una base de datos a partir de un backup bajo determinadas circunstancias, y se suele utilizar con scripts que realizan tareas como la clonación de bases de datos.

1. La salida del siguiente comando se utiliza para recrear los controlfiles para la base de datos migrada.

```
SQL> alter database backup controlfile to trace as '/tmp/waffle.ctrl';
Database altered.
```

2. Después de crear los archivos de control, copie el archivo en el nuevo servidor.

```
[oracle@jfsc3 tmp]$ scp oracle@jfsc1:/tmp/waffle.ctrl /tmp/
oracle@jfsc1's password:
waffle.ctrl                                           100% 5199
5.1KB/s   00:00
```

Archivo de parámetros de copia de seguridad

También se necesita un archivo de parámetros en el nuevo entorno. El método más simple es crear un pfile a partir del spfile o pfile actual. En este ejemplo, la base de datos de origen está utilizando un spfile.

```
SQL> create pfile='/tmp/waffle.tmp.pfile' from spfile;
File created.
```

Crear entrada oratab

La creación de una entrada oratab es necesaria para el correcto funcionamiento de utilidades como oraenv. Para crear una entrada de oratab, realice el siguiente paso.

```
WAFFLE:/orabin/product/12.1.0/dbhome_1:N
```

Preparar la estructura de directorios

Si los directorios necesarios no estaban presentes, debe crearlos o el procedimiento de inicio de la base de datos falla. Para preparar la estructura de directorios, complete los siguientes requisitos mínimos.

```
[oracle@jfsc3 ~]$ . oraenv
ORACLE_SID = [oracle] ? WAFFLE
The Oracle base has been set to /orabin
[oracle@jfsc3 ~]$ cd $ORACLE_BASE
[oracle@jfsc3 orabin]$ cd admin
[oracle@jfsc3 admin]$ mkdir WAFFLE
[oracle@jfsc3 admin]$ cd WAFFLE
[oracle@jfsc3 WAFFLE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Actualizaciones de archivos de parámetros

1. Para copiar el archivo de parámetros en el nuevo servidor, ejecute los siguientes comandos. La ubicación predeterminada es la `$ORACLE_HOME/dbs` directorio. En este caso, el archivo pfile se puede colocar en cualquier lugar. Sólo se utiliza como paso intermedio en el proceso de migración.

```
[oracle@jfsc3 admin]$ scp oracle@jfsc1:/tmp/waffle.tmp.pfile
$ORACLE_HOME/dbs/waffle.tmp.pfile
oracle@jfsc1's password:
waffle.pfile                                          100%  916
0.9KB/s   00:00
```

1. Edite el archivo según sea necesario. Por ejemplo, si la ubicación del archive log ha cambiado, el archivo pfile debe modificarse para reflejar la nueva ubicación. En este ejemplo, sólo se reubican los archivos de control, en parte para distribuirlos entre los sistemas de archivos de registro y de datos.

```
[root@jfsc1 tmp]# cat waffle.pfile
WAFFLE.__data_transfer_cache_size=0
WAFFLE.__db_cache_size=507510784
WAFFLE.__java_pool_size=4194304
WAFFLE.__large_pool_size=20971520
WAFFLE.__oracle_base='/orabin'#ORACLE_BASE set from environment
WAFFLE.__pga_aggregate_target=268435456
WAFFLE.__sga_target=805306368
WAFFLE.__shared_io_pool_size=29360128
WAFFLE.__shared_pool_size=234881024
WAFFLE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/WAFFLE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/oradata//WAFFLE/control01.ctl','/oradata//WAFFLE/contr
ol02.ctl'
*.control_files='/oradata/WAFFLE/control01.ctl','/logs/WAFFLE/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='WAFFLE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=WAFFLEXDB)'
*.log_archive_dest_1='LOCATION=/logs/WAFFLE/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

2. Una vez finalizadas las ediciones, cree un archivo spfile basado en este archivo pfile.

```
SQL> create spfile from pfile='waffle.tmp.pfile';
File created.
```

Vuelva a crear los archivos de control

En un paso anterior, la salida de `backup controlfile to trace` se ha copiado en el nuevo servidor. La parte específica de la salida necesaria es la `controlfile recreation` comando. Esta información se puede encontrar en el archivo bajo la sección marcada `Set #1. NORESETLOGS`. Comienza con la línea `create controlfile reuse database` y debe incluir la palabra `noresetlogs`. Termina con el carácter de punto y coma (;).

1. En este procedimiento de ejemplo, el archivo se lee de la siguiente manera.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/logs/WAFFLE/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/logs/WAFFLE/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/logs/WAFFLE/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

2. Edite este script como desee para reflejar la nueva ubicación de los distintos archivos. Por ejemplo, algunos archivos de datos conocidos por admitir una gran I/O podrían redirigirse a un sistema de archivos en un nivel de almacenamiento de alto rendimiento. En otros casos, los cambios podrían ser únicamente por motivos de administrador, como el aislamiento de los archivos de datos de una PDB determinada en volúmenes dedicados.
3. En este ejemplo, la `DATAFILE` stanza se deja sin cambios, pero los redo logs se mueven a una nueva ubicación en `/redo` en lugar de compartir espacio con archive logs `/logs`.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
SQL> CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
  2      MAXLOGFILES 16
  3      MAXLOGMEMBERS 3
  4      MAXDATAFILES 100
  5      MAXINSTANCES 8
  6      MAXLOGHISTORY 292
  7  LOGFILE
  8    GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  9    GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
 10    GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
 11  -- STANDBY LOGFILE
 12  DATAFILE
 13    '/oradata/WAFFLE/system01.dbf',
 14    '/oradata/WAFFLE/sysaux01.dbf',
 15    '/oradata/WAFFLE/undotbs01.dbf',
 16    '/oradata/WAFFLE/users01.dbf'
 17  CHARACTER SET WE8MSWIN1252
 18  ;
Control file created.
SQL>
```

Si alguno de los archivos está mal ubicado o los parámetros están mal configurados, se generan errores que indican lo que debe corregirse. La base de datos está montada, pero aún no está abierta y no se puede abrir porque los archivos de datos en uso siguen marcados como en modo de copia de seguridad en caliente. Los archive logs deben aplicarse primero para que la base de datos sea coherente.

Replicación de registro inicial

Se necesita al menos una operación de respuesta de log para que los archivos de datos sean consistentes. Hay muchas opciones disponibles para reproducir logs. En algunos casos, la ubicación original del archive log en el servidor original se puede compartir a través de NFS, y la respuesta del log se puede realizar directamente. En otros casos, los archive logs deben copiarse.

Por ejemplo, un simple `scp` la operación puede copiar todos los registros actuales del servidor de origen al servidor de migración:

```
[oracle@jfsc3 arch]$ scp jfsc1:/logs/WAFFLE/arch/* ./
oracle@jfsc1's password:
1_22_912662036.dbf                                          100%   47MB
47.0MB/s   00:01
1_23_912662036.dbf                                          100%   40MB
40.4MB/s   00:00
1_24_912662036.dbf                                          100%   45MB
45.4MB/s   00:00
1_25_912662036.dbf                                          100%   41MB
40.9MB/s   00:01
1_26_912662036.dbf                                          100%   39MB
39.4MB/s   00:00
1_27_912662036.dbf                                          100%   39MB
38.7MB/s   00:00
1_28_912662036.dbf                                          100%   40MB
40.1MB/s   00:01
1_29_912662036.dbf                                          100%   17MB
16.9MB/s   00:00
1_30_912662036.dbf                                          100%  636KB
636.0KB/s   00:00
```

Reproducción de log inicial

Una vez que los archivos están en la ubicación del archive log, se pueden reproducir emitiendo el comando `recover database until cancel` seguido de la respuesta `AUTO` para reproducir automáticamente todos los logs disponibles.

```
SQL> recover database until cancel;
ORA-00279: change 382713 generated at 05/24/2016 09:00:54 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_23_912662036.dbf
ORA-00280: change 382713 for thread 1 is in sequence #23
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 405712 generated at 05/24/2016 15:01:05 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_24_912662036.dbf
ORA-00280: change 405712 for thread 1 is in sequence #24
ORA-00278: log file '/logs/WAFFLE/arch/1_23_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
ORA-00278: log file '/logs/WAFFLE/arch/1_30_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_31_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

La respuesta final del archive log informa de un error, pero esto es normal. El registro lo indica `sqlplus` estaba buscando un archivo de registro en particular y no lo encontró. La razón es, lo más probable, que el archivo log no existe aún.

Si la base de datos de origen se puede cerrar antes de copiar archive logs, este paso debe realizarse una sola vez. Los archive logs se copian y se reproducen y, a continuación, el proceso puede continuar directamente con el proceso de transposición que replica los redo logs críticos.

Replicación y repetición de log incremental

En la mayoría de los casos, la migración no se realiza de forma inmediata. Pueden pasar días o incluso semanas antes de que se complete el proceso de migración, lo que significa que los registros deben enviarse continuamente a la base de datos de réplica y reproducirse. Por lo tanto, al llegar la transición, es necesario transferir y reproducir unos datos mínimos.

Al hacerlo se puede ejecutar un script de muchas maneras, pero uno de los métodos más populares es usar rsync, una utilidad común de replicación de archivos. La forma más segura de utilizar esta utilidad es configurarla como daemon. Por ejemplo, la `rsyncd.conf` el siguiente archivo muestra cómo crear un recurso llamado `waffle.arch` Al que se accede con las credenciales de usuario de Oracle y se asigna a. `/logs/WAFFLE/arch`. Lo que es más importante, el recurso se establece en solo lectura, lo que permite que los datos de producción se lean, pero no se alteren.

```
[root@jfsc1 arch]# cat /etc/rsyncd.conf
[waffle.arch]
   uid=oracle
   gid=dba
   path=/logs/WAFFLE/arch
   read only = true
[root@jfsc1 arch]# rsync --daemon
```

El siguiente comando sincroniza el destino del archive log del nuevo servidor con el recurso rsync `waffle.arch` en el servidor original. La `t` argumento en `rsync - potg` hace que la lista de archivos se compare en función de la marca de tiempo, y solo se copian los archivos nuevos. Este proceso proporciona una actualización incremental del nuevo servidor. Este comando también se puede programar en cron para que se ejecute de forma regular.

```
[oracle@jfsc3 arch]$ rsync -potg --stats --progress jfsc1::waffle.arch/*
/logs/WAFFLE/arch/
1_31_912662036.dbf
      650240 100%  124.02MB/s    0:00:00 (xfer#1, to-check=8/18)
1_32_912662036.dbf
     4873728 100%  110.67MB/s    0:00:00 (xfer#2, to-check=7/18)
1_33_912662036.dbf
     4088832 100%   50.64MB/s    0:00:00 (xfer#3, to-check=6/18)
1_34_912662036.dbf
     8196096 100%   54.66MB/s    0:00:00 (xfer#4, to-check=5/18)
1_35_912662036.dbf
    19376128 100%   57.75MB/s    0:00:00 (xfer#5, to-check=4/18)
1_36_912662036.dbf
       71680 100%  201.15kB/s    0:00:00 (xfer#6, to-check=3/18)
1_37_912662036.dbf
     1144320 100%    3.06MB/s    0:00:00 (xfer#7, to-check=2/18)
1_38_912662036.dbf
    35757568 100%   63.74MB/s    0:00:00 (xfer#8, to-check=1/18)
1_39_912662036.dbf
      984576 100%    1.63MB/s    0:00:00 (xfer#9, to-check=0/18)
Number of files: 18
Number of files transferred: 9
Total file size: 399653376 bytes
Total transferred file size: 75143168 bytes
Literal data: 75143168 bytes
Matched data: 0 bytes
File list size: 474
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 204
Total bytes received: 75153219
sent 204 bytes  received 75153219 bytes  150306846.00 bytes/sec
total size is 399653376  speedup is 5.32
```

Una vez recibidos los registros, deben reproducirse. Ejemplos anteriores muestran el uso de sqlplus para ejecutar manualmente `recover database until cancel`, un proceso que se puede automatizar fácilmente. El ejemplo que se muestra aquí utiliza el script descrito en "Reproducir Logs en Base de Datos". Los scripts aceptan un argumento que especifica la base de datos que necesita una operación de reproducción. Esto permite utilizar el mismo script en un esfuerzo de migración de varias bases de datos.

```
[oracle@jfsc3 logs]$ ./replay.logs.pl WAFFLE
ORACLE_SID = [WAFFLE] ? The Oracle base remains unchanged with value
/orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu May 26 10:47:16 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 814256 generated at 05/26/2016 04:52:30 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_32_912662036.dbf
ORA-00280: change 814256 for thread 1 is in sequence #32
ORA-00278: log file '/logs/WAFFLE/arch/1_31_912662036.dbf' no longer
needed for
this recovery
ORA-00279: change 814780 generated at 05/26/2016 04:53:04 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_33_912662036.dbf
ORA-00280: change 814780 for thread 1 is in sequence #33
ORA-00278: log file '/logs/WAFFLE/arch/1_32_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
ORA-00278: log file '/logs/WAFFLE/arch/1_39_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Transición

Cuando esté listo para realizar la transición al nuevo entorno, debe realizar una sincronización final que incluya tanto archive logs como redo logs. Si la ubicación de redo log original no se conoce todavía, se puede identificar de la siguiente manera:

```
SQL> select member from v$logfile;
MEMBER
-------------------------------------------------------------------------------
------
/logs/WAFFLE/redo/redo01.log
/logs/WAFFLE/redo/redo02.log
/logs/WAFFLE/redo/redo03.log
```

1. Cierre la base de datos de origen.
2. Realice una sincronización final de los archive logs en el nuevo servidor con el método deseado.
3. Los redo logs de origen se deben copiar en el nuevo servidor. En este ejemplo, los redo logs se reubicaron en un nuevo directorio en `/redo`.

   ```
   [oracle@jfsc3 logs]$ scp jfsc1:/logs/WAFFLE/redo/* /redo/
   oracle@jfsc1's password:
   redo01.log
   100%   50MB  50.0MB/s   00:01
   redo02.log
   100%   50MB  50.0MB/s   00:00
   redo03.log
   100%   50MB  50.0MB/s   00:00
   ```

4. En esta etapa, el nuevo entorno de base de datos contiene todos los archivos necesarios para llevarlo al mismo estado que el origen. Los registros de archivos se deben reproducir por última vez.

```
SQL> recover database until cancel;
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

5. Una vez finalizado, los redo logs se deben volver a reproducir. Si el mensaje `Media recovery complete` se devuelve, el proceso se realiza correctamente y las bases de datos se sincronizan y se pueden abrir.

```
SQL> recover database;
Media recovery complete.
SQL> alter database open;
Database altered.
```

Envío de registros: ASM al sistema de archivos

Este ejemplo muestra el uso de Oracle RMAN para migrar una base de datos. Es muy similar al ejemplo anterior del envío de registros del sistema de archivos al sistema de archivos, pero los archivos de ASM no son visibles para el host. La única opción para migrar datos ubicados en dispositivos ASM es mediante la reubicación del LUN de ASM o mediante Oracle RMAN para realizar las operaciones de copia.

Aunque RMAN es un requisito para copiar archivos de Oracle ASM, el uso de RMAN no se limita a ASM. RMAN se puede utilizar para migrar de cualquier tipo de almacenamiento a cualquier otro tipo.

Este ejemplo muestra la reubicación de una base de datos llamada PANCAKE del almacenamiento de ASM a un sistema de archivos normal ubicado en un servidor diferente en las rutas de acceso `/oradata` y.. `/logs`.

Crear copia de seguridad de base de datos

El primer paso es crear una copia de seguridad de la base de datos que se migrará a un servidor alternativo. Dado que el origen utiliza Oracle ASM, se debe utilizar RMAN. Se puede realizar una copia de seguridad simple de RMAN del siguiente modo. Este método crea una copia de seguridad etiquetada que RMAN puede identificar fácilmente más adelante en el procedimiento.

El primer comando define el tipo de destino para la copia de seguridad y la ubicación que se utilizará. El segundo inicia la copia de seguridad de los archivos de datos solamente.

```
RMAN> configure channel device type disk format '/rman/pancake/%U';
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters are successfully stored
RMAN> backup database tag 'ONTAP_MIGRATION';
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=251 device type=DISK
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/PANCAKE/system01.dbf
input datafile file number=00002 name=+ASM0/PANCAKE/sysaux01.dbf
input datafile file number=00003 name=+ASM0/PANCAKE/undotbs101.dbf
input datafile file number=00004 name=+ASM0/PANCAKE/users01.dbf
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1gr6c161_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:03
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1hr6c164_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Copia de seguridad del archivo de control

Se necesita un archivo de control de copia de seguridad más adelante en el procedimiento del `duplicate database` funcionamiento.

```
RMAN> backup current controlfile format '/rman/pancake/ctrl.bkp';
Starting backup at 24-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/ctrl.bkp tag=TAG20160524T032651 comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Archivo de parámetros de copia de seguridad

También se necesita un archivo de parámetros en el nuevo entorno. El método más simple es crear un pfile a partir del spfile o pfile actual. En este ejemplo, la base de datos de origen utiliza un spfile.

```
RMAN> create pfile='/rman/pancake/pfile' from spfile;
Statement processed
```

Script de cambio de nombre de archivo de ASM

Varias ubicaciones de archivos definidas actualmente en los controlfiles cambian cuando se mueve la base de datos. El siguiente archivo de comandos crea un archivo de comandos de RMAN para facilitar el proceso. Este ejemplo muestra una base de datos con un número muy pequeño de archivos de datos, pero normalmente las bases de datos contienen cientos o incluso miles de archivos de datos.

Este script se puede encontrar en "Conversión de ASM a Nombre de Sistema de Archivos" y hace dos cosas.

En primer lugar, crea un parámetro para redefinir las ubicaciones de redo log llamadas `log_file_name_convert`. Es esencialmente una lista de campos alternos. El primer campo es la ubicación de un redo log actual y el segundo campo es la ubicación del nuevo servidor. El patrón se repite entonces.

La segunda función consiste en proporcionar una plantilla para el cambio de nombre del archivo de datos. El archivo de comandos pasa por los archivos de datos, extrae la información del nombre y el número de archivo y lo formatea como un archivo de comandos de RMAN. A continuación, hace lo mismo con los archivos temporales. El resultado es un script de rman simple que se puede editar como se desee para asegurarse de que los archivos se restauran en la ubicación deseada.

```
SQL> @/rman/mk.rename.scripts.sql
Parameters for log file conversion:
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/NEW_PATH/redo01.log','+ASM0/PANCAKE/redo02.log',
'/NEW_PATH/redo02.log','+ASM0/PANCAKE/redo03.log', '/NEW_PATH/redo03.log'
rman duplication script:
run
{
set newname for datafile 1 to '+ASM0/PANCAKE/system01.dbf';
set newname for datafile 2 to '+ASM0/PANCAKE/sysaux01.dbf';
set newname for datafile 3 to '+ASM0/PANCAKE/undotbs101.dbf';
set newname for datafile 4 to '+ASM0/PANCAKE/users01.dbf';
set newname for tempfile 1 to '+ASM0/PANCAKE/temp01.dbf';
duplicate target database for standby backup location INSERT_PATH_HERE;
}
PL/SQL procedure successfully completed.
```

Captura la salida de esta pantalla. La `log_file_name_convert` el parámetro se coloca en el archivo pfile como se describe a continuación. El archivo de datos RENAME y el archivo de comandos DUPLICATE de RMAN se deben editar en consecuencia para colocar los archivos de datos en las ubicaciones deseadas. En este ejemplo, se colocan todos `/oradata/pancake`.

```
run
{
set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
set newname for datafile 4 to '/oradata/pancake/users.dbf';
set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
duplicate target database for standby backup location '/rman/pancake';
}
```

Preparar la estructura de directorios

Los scripts están casi listos para ejecutarse, pero primero debe estar la estructura de directorios en su lugar. Si los directorios necesarios no están ya presentes, se deben crear o el procedimiento de inicio de la base de datos falla. El ejemplo siguiente refleja los requisitos mínimos.

```
[oracle@jfsc2 ~]$ mkdir /oradata/pancake
[oracle@jfsc2 ~]$ mkdir /logs/pancake
[oracle@jfsc2 ~]$ cd /orabin/admin
[oracle@jfsc2 admin]$ mkdir PANCAKE
[oracle@jfsc2 admin]$ cd PANCAKE
[oracle@jfsc2 PANCAKE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Crear entrada oratab

El siguiente comando es necesario para que utilidades como oraenv funcionen correctamente.

```
PANCAKE:/orabin/product/12.1.0/dbhome_1:N
```

Actualizaciones de parámetros

El archivo pfile guardado se debe actualizar para reflejar cualquier cambio de ruta en el nuevo servidor. El script de duplicación de RMAN modifica los cambios de la ruta de acceso del archivo de datos y casi todas las bases de datos requieren cambios en el `control_files` y.. `log_archive_dest` parámetros. Es posible que también haya ubicaciones de archivos de auditoría que deban modificarse y parámetros como `db_create_file_dest` Puede que no sea relevante fuera de ASM. Un DBA con experiencia debe revisar cuidadosamente los cambios propuestos antes de continuar.

En este ejemplo, los cambios clave son las ubicaciones del archivo de control, el destino del archivo de registro y la adición del `log_file_name_convert` parámetro.

```
PANCAKE.__data_transfer_cache_size=0
PANCAKE.__db_cache_size=545259520
PANCAKE.__java_pool_size=4194304
PANCAKE.__large_pool_size=25165824
PANCAKE.__oracle_base='/orabin'#ORACLE_BASE set from environment
PANCAKE.__pga_aggregate_target=268435456
PANCAKE.__sga_target=805306368
PANCAKE.__shared_io_pool_size=29360128
PANCAKE.__shared_pool_size=192937984
PANCAKE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/PANCAKE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='+ASM0/PANCAKE/control01.ctl','+ASM0/PANCAKE/control02.ctl
'
*.control_files='/oradata/pancake/control01.ctl','/logs/pancake/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='PANCAKE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=PANCAKEXDB)'
*.log_archive_dest_1='LOCATION=+ASM1'
*.log_archive_dest_1='LOCATION=/logs/pancake'
*.log_archive_format='%t_%s_%r.dbf'
'/logs/path/redo02.log'
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/logs/pancake/redo01.log', '+ASM0/PANCAKE/redo02.log',
'/logs/pancake/redo02.log', '+ASM0/PANCAKE/redo03.log',
'/logs/pancake/redo03.log'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Después de confirmar los nuevos parámetros, los parámetros deben ponerse en vigor. Existen varias opciones, pero la mayoría de los clientes crean un spfile basado en el archivo pfile de texto.

```
bash-4.1$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0 Production on Fri Jan 8 11:17:40 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> create spfile from pfile='/rman/pancake/pfile';
File created.
```

Inicio nomount

El último paso antes de replicar la base de datos es abrir los procesos de la base de datos pero no montar los archivos. En este paso, los problemas con el spfile pueden hacerse evidentes. Si la `startup nomount` el comando falla debido a un error de parámetro, es fácil de cerrar, corregir la plantilla pfile, recargarla como spfile e intentarlo de nuevo.

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             373296240 bytes
Database Buffers          423624704 bytes
Redo Buffers                5455872 bytes
```

Duplique la base de datos

La restauración de la copia de seguridad de RMAN anterior en la nueva ubicación consume más tiempo que otros pasos de este proceso. La base de datos se debe duplicar sin cambiar el identificador de base de datos (DBID) ni restablecer los logs. Esto evita que se apliquen los logs, lo que es un paso necesario para sincronizar completamente las copias.

Conéctese a la base de datos con RMAN como aux y emita el comando DUPLICATE DATABASE mediante el script creado en un paso anterior.

```
[oracle@jfsc2 pancake]$ rman auxiliary /
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 03:04:56
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to auxiliary database: PANCAKE (not mounted)
RMAN> run
2> {
3> set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
4> set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
5> set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
6> set newname for datafile 4 to '/oradata/pancake/users.dbf';
7> set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
```

```
8> duplicate target database for standby backup location '/rman/pancake';
9> }
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting Duplicate Db at 24-MAY-16
contents of Memory Script:
{
   restore clone standby controlfile from  '/rman/pancake/ctrl.bkp';
}
executing Memory Script
Starting restore at 24-MAY-16
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
channel ORA_AUX_DISK_1: restoring control file
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:01
output file name=/oradata/pancake/control01.ctl
output file name=/logs/pancake/control02.ctl
Finished restore at 24-MAY-16
contents of Memory Script:
{
   sql clone 'alter database mount standby database';
}
executing Memory Script
sql statement: alter database mount standby database
released channel: ORA_AUX_DISK_1
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
contents of Memory Script:
{
   set newname for tempfile  1 to
 "/oradata/pancake/temp.dbf";
   switch clone tempfile all;
   set newname for datafile  1 to
 "/oradata/pancake/pancake.dbf";
   set newname for datafile  2 to
 "/oradata/pancake/sysaux.dbf";
   set newname for datafile  3 to
 "/oradata/pancake/undotbs1.dbf";
   set newname for datafile  4 to
 "/oradata/pancake/users.dbf";
   restore
   clone database
   ;
```

```
}
executing Memory Script
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/pancake/temp.dbf in control file
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting restore at 24-MAY-16
using channel ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: starting datafile backup set restore
channel ORA_AUX_DISK_1: specifying datafile(s) to restore from backup set
channel ORA_AUX_DISK_1: restoring datafile 00001 to
/oradata/pancake/pancake.dbf
channel ORA_AUX_DISK_1: restoring datafile 00002 to
/oradata/pancake/sysaux.dbf
channel ORA_AUX_DISK_1: restoring datafile 00003 to
/oradata/pancake/undotbs1.dbf
channel ORA_AUX_DISK_1: restoring datafile 00004 to
/oradata/pancake/users.dbf
channel ORA_AUX_DISK_1: reading from backup piece
/rman/pancake/1gr6c161_1_1
channel ORA_AUX_DISK_1: piece handle=/rman/pancake/1gr6c161_1_1
tag=ONTAP_MIGRATION
channel ORA_AUX_DISK_1: restored backup piece 1
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:07
Finished restore at 24-MAY-16
contents of Memory Script:
{
   switch clone datafile all;
}
executing Memory Script
datafile 1 switched to datafile copy
input datafile copy RECID=5 STAMP=912655725 file
name=/oradata/pancake/pancake.dbf
datafile 2 switched to datafile copy
input datafile copy RECID=6 STAMP=912655725 file
name=/oradata/pancake/sysaux.dbf
datafile 3 switched to datafile copy
input datafile copy RECID=7 STAMP=912655725 file
name=/oradata/pancake/undotbs1.dbf
datafile 4 switched to datafile copy
input datafile copy RECID=8 STAMP=912655725 file
name=/oradata/pancake/users.dbf
Finished Duplicate Db at 24-MAY-16
```

Replicación de registro inicial

Ahora debe enviar los cambios de la base de datos de origen a una nueva ubicación. Si lo hace, puede que sea necesario realizar una combinación de pasos. El método más sencillo sería tener RMAN en la base de datos de origen escribir archive logs en una conexión de red compartida. Si una ubicación compartida no está disponible, un método alternativo es utilizar RMAN para escribir en un sistema de archivos local y, a continuación, utilizar rcp o rsync para copiar los archivos.

En este ejemplo, la `/rman` Directory es un recurso compartido NFS que está disponible tanto para la base de datos original como para la migrada.

Una cuestión importante aquí es la `disk format` cláusula. El formato de disco del backup es `%h_%e_%a.dbf`, Lo que significa que debe utilizar el formato de número de hilo, número de secuencia e identificador de activación para la base de datos. Aunque las letras son diferentes, esto coincide con `log_archive_format='%t_%s_%r.dbf` en el pfile. Este parámetro también especifica archive logs en el formato de Núm. De thread, Núm. De secuencia e ID de activación. El resultado final es que los backups de los archivos de registro del origen utilizan una convención de nomenclatura que espera la base de datos. Al hacerlo, se realizan operaciones como `recover database` mucho más sencillo porque sqlplus anticipa correctamente los nombres de los archive logs que se van a reproducir.

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/arch/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
released channel: ORA_DISK_1
RMAN> backup as copy archivelog from time 'sysdate-2';
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=373 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=70 STAMP=912658508
output file name=/rman/pancake/logship/1_54_912576125.dbf RECID=123
STAMP=912659482
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=29 STAMP=912654101
output file name=/rman/pancake/logship/1_41_912576125.dbf RECID=124
STAMP=912659483
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=33 STAMP=912654688
output file name=/rman/pancake/logship/1_45_912576125.dbf RECID=152
STAMP=912659514
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=36 STAMP=912654809
output file name=/rman/pancake/logship/1_47_912576125.dbf RECID=153
STAMP=912659515
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Reproducción de log inicial

Una vez que los archivos están en la ubicación del archive log, se pueden reproducir emitiendo el comando `recover database until cancel` seguido de la respuesta `AUTO` para reproducir automáticamente todos los logs disponibles. El archivo de parámetros está dirigiendo los archive logs al `/logs/archive`, Pero esto no coincide con la ubicación en la que se utilizó RMAN para guardar registros. La ubicación se puede redirigir temporalmente de la siguiente manera antes de recuperar la base de datos.

```
SQL> alter system set log_archive_dest_1='LOCATION=/rman/pancake/logship'
scope=memory;
System altered.
SQL> recover standby database until cancel;
ORA-00279: change 560224 generated at 05/24/2016 03:25:53 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_49_912576125.dbf
ORA-00280: change 560224 for thread 1 is in sequence #49
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 560353 generated at 05/24/2016 03:29:17 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_50_912576125.dbf
ORA-00280: change 560353 for thread 1 is in sequence #50
ORA-00278: log file '/rman/pancake/logship/1_49_912576125.dbf' no longer
needed
for this recovery
...
ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
ORA-00278: log file '/rman/pancake/logship/1_53_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_54_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

La respuesta final del archive log informa de un error, pero esto es normal. El error indica que sqlplus estaba buscando un archivo log en particular y no lo encontró. La razón es más probable que el archivo log no exista aún.

Si la base de datos de origen se puede cerrar antes de copiar archive logs, este paso debe realizarse una sola vez. Los archive logs se copian y se reproducen y, a continuación, el proceso puede continuar directamente con el proceso de transposición que replica los redo logs críticos.

Replicación y repetición de log incremental

En la mayoría de los casos, la migración no se realiza de forma inmediata. Pueden pasar días o incluso semanas antes de que se complete el proceso de migración, lo que significa que los registros deben enviarse continuamente a la base de datos de réplica y reproducirse. Al hacerlo, se garantiza que se deban transferir y reproducir unos datos mínimos al llegar la transición.

Este proceso se puede programar fácilmente. Por ejemplo, el siguiente comando se puede programar en la base de datos original para asegurarse de que la ubicación utilizada para el envío de registros se actualiza

continuamente.

```
[oracle@jfsc1 pancake]$ cat copylogs.rman
configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
backup as copy archivelog from time 'sysdate-2';
```

```
[oracle@jfsc1 pancake]$ rman target / cmdfile=copylogs.rman
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 04:36:19
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: PANCAKE (DBID=3574534589)
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=369 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:22
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=124 STAMP=912659483
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:23
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_41_912576125.dbf
continuing other job steps, job failed will not be re-run
...
```

```
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:55
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:57
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
Recovery Manager complete.
```

Una vez recibidos los registros, deben reproducirse. Ejemplos anteriores mostraron el uso de sqlplus para ejecutar manualmente `recover database until cancel`, que se puede automatizar fácilmente. El ejemplo que se muestra aquí utiliza el script descrito en "Logs de Reproducción en Base de Datos en Espera". El script acepta un argumento que especifica la base de datos que necesita una operación de reproducción. Este proceso permite utilizar el mismo script en un esfuerzo de migración de varias bases de datos.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 04:47:10 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 562219 generated at 05/24/2016 04:15:08 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_55_912576125.dbf
ORA-00280: change 562219 for thread 1 is in sequence #55
ORA-00278: log file '/rman/pancake/logship/1_54_912576125.dbf' no longer
needed for this recovery
ORA-00279: change 562370 generated at 05/24/2016 04:19:18 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_56_912576125.dbf
ORA-00280: change 562370 for thread 1 is in sequence #56
ORA-00278: log file '/rman/pancake/logship/1_55_912576125.dbf' no longer
needed for this recovery
...
ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
ORA-00278: log file '/rman/pancake/logship/1_64_912576125.dbf' no longer
needed for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_65_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Transición

Cuando esté listo para pasar al nuevo entorno, debe realizar una sincronización final. Cuando se trabaja con sistemas de archivos normales, es fácil asegurarse de que la base de datos migrada esté sincronizada al 100% con la original, ya que los redo logs originales se copian y se vuelven a reproducir. No hay una buena forma de hacerlo con ASM. Sólo los archive logs se pueden volver a copiar fácilmente. Para asegurarse de que no se pierden datos, el cierre final de la base de datos original debe realizarse con cuidado.

1. En primer lugar, la base de datos debe estar en modo inactivo, asegurándose de que no se realicen cambios. Esta desactivación puede incluir la desactivación de las operaciones programadas, el cierre de listeners y/o el cierre de aplicaciones.
2. Después de realizar este paso, la mayoría de los DBA crean una tabla ficticia que sirve como marcador del cierre.
3. Forzar un archivo log para asegurarse de que la creación de la tabla ficticia se registra en los archive logs. Para ello, ejecute los siguientes comandos:

```
SQL> create table cutovercheck as select * from dba_users;
Table created.
SQL> alter system archive log current;
System altered.
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
```

4. Para copiar el último de los archive logs, ejecute los siguientes comandos. La base de datos debe estar disponible pero no abierta.

```
SQL> startup mount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
Database mounted.
```

5. Para copiar los archive logs, ejecute los siguientes comandos:

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=8 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:24
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:58
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:59:00
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
```

6. Por último, vuelva a reproducir los archive logs restantes en el nuevo servidor.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 05:00:53 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed
for thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 563629 generated at 05/24/2016 04:55:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_66_912576125.dbf
ORA-00280: change 563629 for thread 1 is in sequence #66
ORA-00278: log file '/rman/pancake/logship/1_65_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_66_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

7. En esta fase, replique todos los datos. La base de datos está lista para convertirse de una base de datos en espera a una base de datos operativa activa y, a continuación, abrirse.

```
SQL> alter database activate standby database;
Database altered.
SQL> alter database open;
Database altered.
```

8. Confirme la presencia de la tabla ficticia y, a continuación, suéltela.

```
SQL> desc cutovercheck
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 USERNAME                                  NOT NULL VARCHAR2(128)
 USER_ID                                   NOT NULL NUMBER
 PASSWORD                                           VARCHAR2(4000)
 ACCOUNT_STATUS                            NOT NULL VARCHAR2(32)
 LOCK_DATE                                          DATE
 EXPIRY_DATE                                        DATE
 DEFAULT_TABLESPACE                        NOT NULL VARCHAR2(30)
 TEMPORARY_TABLESPACE                      NOT NULL VARCHAR2(30)
 CREATED                                   NOT NULL DATE
 PROFILE                                   NOT NULL VARCHAR2(128)
 INITIAL_RSRC_CONSUMER_GROUP                        VARCHAR2(128)
 EXTERNAL_NAME                                      VARCHAR2(4000)
 PASSWORD_VERSIONS                                  VARCHAR2(12)
 EDITIONS_ENABLED                                   VARCHAR2(1)
 AUTHENTICATION_TYPE                                VARCHAR2(8)
 PROXY_ONLY_CONNECT                                 VARCHAR2(1)
 COMMON                                             VARCHAR2(3)
 LAST_LOGIN                                         TIMESTAMP(9) WITH
TIME ZONE
 ORACLE_MAINTAINED                                  VARCHAR2(1)
SQL> drop table cutovercheck;
Table dropped.
```

Migración de redo log no disruptiva

Hay veces en las que una base de datos está correctamente organizada en general con la excepción de los redo logs. Esto puede ocurrir por muchos motivos, el más común de los cuales está relacionado con las copias Snapshot. Productos como SnapManager para Oracle, SnapCenter y el marco de gestión de almacenamiento Snap Creator de NetApp permiten la recuperación casi instantánea de una base de datos, pero únicamente si revierte el estado de los volúmenes de archivos de datos. Si los redo logs comparten espacio con los archivos de datos, la reversión no se puede realizar de forma segura porque podría provocar la destrucción de los redo logs, lo que probablemente significa pérdida de datos. Por lo tanto, los redo logs deben reubicarse.

Este procedimiento es sencillo y puede realizarse sin interrupciones.

Configuración actual de redo log

1. Identifique el Núm. De grupos de redo logs y sus respectivos Núm.s de grupo.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
rows selected.
```

2. Introduzca el tamaño de los redo logs.

```
SQL> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
--------------------------------------------------------------------------------
--------
1 524288000
2 524288000
3 524288000
```

Crear nuevos logs

1. Para cada redo log, cree un nuevo grupo con un tamaño y un Núm. De miembros coincidentes.

```
SQL> alter database add logfile ('/newredo0/redo01a.log',
'/newredo1/redo01b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo02a.log',
'/newredo1/redo02b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo03a.log',
'/newredo1/redo03b.log') size 500M;
Database altered.
SQL>
```

2. Verifique la nueva configuración.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
4 /newredo0/redo01a.log
4 /newredo1/redo01b.log
5 /newredo0/redo02a.log
5 /newredo1/redo02b.log
6 /newredo0/redo03a.log
6 /newredo1/redo03b.log
12 rows selected.
```

Borre los registros antiguos

1. Borre los registros antiguos (grupos 1, 2 y 3).

```
SQL> alter database drop logfile group 1;
Database altered.
SQL> alter database drop logfile group 2;
Database altered.
SQL> alter database drop logfile group 3;
Database altered.
```

2. Si encuentra un error que le impide borrar un log activo, fuerce un cambio al siguiente log para liberar el bloqueo y forzar un punto de control global. Vea el siguiente ejemplo de este proceso. Se ha denegado el intento de borrar el grupo de archivos de registro 2, que se encontraba en la ubicación anterior, porque todavía había datos activos en este archivo de registro.

```
SQL> alter database drop logfile group 2;
alter database drop logfile group 2
*
ERROR at line 1:
ORA-01623: log 2 is current log for instance NTAP (thread 1) - cannot
drop
ORA-00312: online log 2 thread 1: '/redo0/NTAP/redo02a.log'
ORA-00312: online log 2 thread 1: '/redo1/NTAP/redo02b.log'
```

3. Un archivo log seguido de un punto de control permite borrar el archivo log.

```
SQL> alter system archive log current;
System altered.
SQL> alter system checkpoint;
System altered.
SQL> alter database drop logfile group 2;
Database altered.
```

4. A continuación, elimine los registros del sistema de archivos. Debe realizar este proceso con extremo cuidado.

Copiado de datos de host

Al igual que sucede con la migración a nivel de base de datos, la migración en la capa de host proporciona un enfoque independiente del proveedor de almacenamiento.

En otras palabras, en algún momento “solo copiar los archivos” es la mejor opción.

Aunque este enfoque de baja tecnología puede parecer demasiado básico, ofrece beneficios significativos porque no se requiere ningún software especial y los datos originales permanecen intactos de forma segura durante el proceso. La principal limitación es el hecho de que una migración de datos de copia de archivos es un proceso disruptivo, ya que la base de datos debe cerrarse antes de que comience la operación de copia. No hay una buena manera de sincronizar los cambios dentro de un archivo, por lo que los archivos deben estar completamente desactivados antes de que comience la copia.

Si el cierre necesario para una operación de copia no es deseable, la siguiente mejor opción basada en host es utilizar un gestor de volúmenes lógicos (LVM). Existen muchas opciones de LVM, incluido Oracle ASM, todas con capacidades similares, pero también con algunas limitaciones que deben tenerse en cuenta. En la mayoría de los casos, la migración se puede realizar sin tiempos de inactividad ni interrupciones.

Copiando sistema de archivos al sistema de archivos

La utilidad de una operación de copia simple no debe subestimarse. Esta operación requiere un tiempo de inactividad durante el proceso de copia, pero es un proceso muy fiable y no requiere experiencia especial en sistemas operativos, bases de datos o sistemas de almacenamiento. Además, es muy seguro porque no afecta a los datos originales. Normalmente, un administrador de sistemas cambia los sistemas de archivos de origen para montarse como de solo lectura y luego reinicia un servidor para garantizar que nada pueda dañar los datos actuales. El proceso de copia se puede programar para asegurarse de que se ejecuta lo más rápido posible sin riesgo de error por parte del usuario. Dado que el tipo de I/O es una transferencia secuencial simple de datos, es altamente eficiente del ancho de banda.

El siguiente ejemplo muestra una opción para una migración segura y rápida.

Entorno Oracle

El entorno que se va a migrar es el siguiente:

- Sistemas de archivos actuales

```
ontap-nfs1:/host1_oradata       52428800  16196928  36231872  31%
/oradata
ontap-nfs1:/host1_logs          49807360    548032  49259328   2% /logs
```

- Sistemas de archivos nuevos

```
ontap-nfs1:/host1_logs_new      49807360       128  49807232   1%
/new/logs
ontap-nfs1:/host1_oradata_new   49807360       128  49807232   1%
/new/oradata
```

Descripción general

El DBA puede migrar la base de datos simplemente cerrando la base de datos y copiando los archivos, pero el proceso se ejecuta fácilmente en la secuencia de comandos si se deben migrar muchas bases de datos o si se minimiza el tiempo de inactividad es crítico. El uso de scripts también reduce la posibilidad de errores de los usuarios.

Los scripts de ejemplo que se muestran automatizan las siguientes operaciones:

- Cerrando la base de datos
- Convertir los sistemas de archivos existentes a un estado de sólo lectura
- Copia de todos los datos de los sistemas de archivos de origen a los de destino, lo que conserva todos los permisos de archivos
- Desmontaje de los sistemas de archivos antiguos y nuevos
- Volver a montar los nuevos sistemas de archivos en las mismas rutas que los sistemas de archivos anteriores

Procedimiento

1. Cierre la base de datos.

```
[root@host1 current]# ./dbshut.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 15:58:48 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
```

2. Convierta los sistemas de archivos a sólo lectura. Esto se puede hacer más rápidamente usando un script, como se muestra en "Convertir sistema de archivos a Sólo lectura".

```
[root@host1 current]# ./mk.fs.readonly.pl /oradata
/oradata unmounted
/oradata mounted read-only
[root@host1 current]# ./mk.fs.readonly.pl /logs
/logs unmounted
/logs mounted read-only
```

3. Confirme que los sistemas de archivos ahora son de sólo lectura.

```
ontap-nfs1:/host1_oradata on /oradata type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_logs on /logs type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

4. Sincronice el contenido del sistema de archivos con `rsync` comando.

```
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /oradata/ /new/oradata/
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
```

```
 10737426432 100%  153.50MB/s    0:01:06 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%   12.09MB/s    0:00:01 (xfer#2, to-check=9/13)
...
NTAP/undotbs02.dbf
  1073750016 100%  131.60MB/s    0:00:07 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    3.95MB/s    0:00:01 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  162204017.96 bytes/sec
total size is 18570092218  speedup is 1.00
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /logs/ /new/logs/
sending incremental file list
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%   95.98MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   49.45MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   44.68MB/s    0:00:01 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   68.03MB/s    0:00:00 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
```

```
sent 527098156 bytes  received 278 bytes  95836078.91 bytes/sec
total size is 527032832  speedup is 1.00
```

5. Desmonte los sistemas de archivos antiguos y reubique los datos copiados. Esto se puede hacer más rápidamente usando un script, como se muestra en "Reemplazar sistema de archivos".

```
[root@host1 current]# ./swap.fs.pl /logs,/new/logs
/new/logs unmounted
/logs unmounted
Updated /logs mounted
[root@host1 current]# ./swap.fs.pl /oradata,/new/oradata
/new/oradata unmounted
/oradata unmounted
Updated /oradata mounted
```

6. Confirme que los nuevos sistemas de archivos están en posición.

```
ontap-nfs1:/host1_logs_new on /logs type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_oradata_new on /oradata type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

7. Inicie la base de datos.

```
[root@host1 current]# ./dbstart.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 16:10:07 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
```

Transición totalmente automatizada

Este script de ejemplo acepta argumentos del SID de la base de datos seguidos de pares de sistemas de archivos delimitados comúnmente. Para el ejemplo mostrado anteriormente, el comando se emite del siguiente modo:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
```

Cuando se ejecuta, el script de ejemplo intenta realizar la siguiente secuencia. Termina si encuentra un error en cualquier paso:

1. Cierre la base de datos.
2. Convierta los sistemas de archivos actuales al estado de sólo lectura.
3. Utilice cada par delimitado por comas de argumentos del sistema de archivos y sincronice el primer sistema de archivos con el segundo.
4. Desmonte los sistemas de archivos anteriores.
5. Actualice el `/etc/fstab` el archivo es el siguiente:
   a. Cree un backup en `/etc/fstab.bak`.
   b. Comente las entradas anteriores de los sistemas de archivos anteriores y nuevos.
   c. Cree una nueva entrada para el nuevo sistema de archivos que utilice el antiguo punto de montaje.
6. Monte los sistemas de archivos.
7. Inicie la base de datos.

El siguiente texto proporciona un ejemplo de ejecución para este script:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:05:50 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
```

```
sending incremental file list
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%  185.40MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   81.34MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   70.42MB/s    0:00:00 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   47.08MB/s    0:00:01 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
sent 527098156 bytes  received 278 bytes  150599552.57 bytes/sec
total size is 527032832  speedup is 1.00
Succesfully replicated filesystem /logs to /new/logs
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
 10737426432 100%  176.55MB/s    0:00:58 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%    9.48MB/s    0:00:02 (xfer#2, to-check=9/13)
... NTAP/undotbs01.dbf
   309338112 100%   70.76MB/s    0:00:04 (xfer#9, to-check=2/13)
NTAP/undotbs02.dbf
  1073750016 100%  187.65MB/s    0:00:05 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    5.09MB/s    0:00:00 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
```

```
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  177725933.55 bytes/sec
total size is 18570092218  speedup is 1.00
Succesfully replicated filesystem /oradata to /new/oradata
swap 0 /logs /new/logs
/new/logs unmounted
/logs unmounted
Mounted updated /logs
Swapped filesystem /logs for /new/logs
swap 1 /oradata /new/oradata
/new/oradata unmounted
/oradata unmounted
Mounted updated /oradata
Swapped filesystem /oradata for /new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:08:59 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
[root@host1 current]#
```

Migración de Oracle ASM spfile y passwd

Una dificultad para completar la migración que implica ASM es el spfile específico de ASM y el archivo de contraseñas. Por defecto, estos archivos de metadatos críticos se crean en el primer grupo de discos de ASM definido. Si se debe evacuar y eliminar un grupo de discos de ASM concreto, se debe reubicar el archivo spfile y de contraseñas que rigen dicha instancia de ASM.

Otro caso de uso en el que es posible que sea necesario reubicar estos archivos es durante un despliegue de software de gestión de base de datos como SnapManager para Oracle o el complemento de Oracle de SnapCenter. Una de las características de estos productos es restaurar rápidamente una base de datos mediante la reversión del estado de las LUN de ASM que alojan los archivos de datos. Para hacerlo, es necesario desconectar el grupo de discos de ASM antes de realizar una restauración. Esto no es un problema

siempre que los archivos de datos de una base de datos determinada estén aislados en un grupo de discos de ASM dedicado.

Cuando ese grupo de discos también contiene el archivo spfile/passwd de ASM, la única forma en que el grupo de discos se puede poner fuera de línea es cerrar toda la instancia de ASM. Este es un proceso disruptivo, lo que significa que el archivo spfile/passwd tendría que ser reubicado.

Entorno Oracle

1. SID de base de datos = TOAST
2. Archivos de datos actuales en `+DATA`
3. Archivos log y archivos de control actuales en `+LOGS`
4. Se han establecido nuevos grupos de discos de ASM como `+NEWDATA` y.. `+NEWLOGS`

Ubicaciones de archivos spfile/passwd de ASM

La reubicación de estos archivos puede realizarse de forma no disruptiva. Sin embargo, por motivos de seguridad, NetApp recomienda cerrar el entorno de la base de datos para que pueda estar seguro de que los archivos se han reubicado y que la configuración se ha actualizado correctamente. Este procedimiento se debe repetir si hay varias instancias de ASM presentes en un servidor.

Identificar instancias de ASM

Identifique las instancias de ASM en función de los datos registrados en la `oratab` archivo. Las instancias de ASM se indican con un símbolo +.

```
-bash-4.1$ cat /etc/oratab | grep '^+'
+ASM:/orabin/grid:N             # line added by Agent
```

Hay una instancia de ASM denominada +ASM en este servidor.

Asegúrese de que todas las bases de datos están cerradas

El único proceso smon visible debe ser smon para la instancia de ASM en uso. La presencia de otro proceso smon indica que una base de datos todavía está en ejecución.

```
-bash-4.1$ ps -ef | grep smon
oracle     857     1  0 18:26 ?        00:00:00 asm_smon_+ASM
```

El único proceso smon es la propia instancia de ASM. Esto significa que no se ejecuta ninguna otra base de datos y es seguro continuar sin riesgo de interrumpir las operaciones de la base de datos.

Localizar archivos

Identifique la ubicación actual del archivo spfile y de contraseña de ASM mediante `spget` y.. `pwget` comandos.

```
bash-4.1$ asmcmd
ASMCMD> spget
+DATA/spfile.ora
```

```
ASMCMD> pwget --asm
+DATA/orapwasm
```

Los archivos se encuentran en la base del `+DATA` grupo de discos.

Copiar archivos

Copie los archivos en el nuevo grupo de discos de ASM con `spcopy` y.. `pwcopy` comandos. Si el nuevo grupo de discos se ha creado recientemente y está vacío actualmente, es posible que tenga que montarlo primero.

```
ASMCMD> mount NEWDATA
```

```
ASMCMD> spcopy +DATA/spfile.ora +NEWDATA/spfile.ora
copying +DATA/spfile.ora -> +NEWDATA/spfilea.ora
```

```
ASMCMD> pwcopy +DATA/orapwasm +NEWDATA/orapwasm
copying +DATA/orapwasm -> +NEWDATA/orapwasm
```

Los archivos se han copiado ahora de `+DATA` para `+NEWDATA`.

Actualizar instancia de ASM

La instancia de ASM debe actualizarse para reflejar el cambio de ubicación. La `spset` y.. `pwset` Los comandos actualizan los metadatos de ASM necesarios para iniciar el grupo de discos de ASM.

```
ASMCMD> spset +NEWDATA/spfile.ora
ASMCMD> pwset --asm +NEWDATA/orapwasm
```

Active ASM con archivos actualizados

En este punto, la instancia de ASM sigue utilizando las ubicaciones anteriores de estos archivos. La instancia se debe reiniciar para forzar una nueva lectura de los archivos desde sus nuevas ubicaciones y liberar bloqueos en los archivos anteriores.

```
-bash-4.1$ sqlplus / as sysasm
SQL> shutdown immediate;
ASM diskgroups volume disabled
ASM diskgroups dismounted
ASM instance shutdown
```

```
SQL> startup
ASM instance started
Total System Global Area 1140850688 bytes
Fixed Size                  2933400 bytes
Variable Size            1112751464 bytes
ASM Cache                  25165824 bytes
ORA-15032: not all alterations performed
ORA-15017: diskgroup "NEWDATA" cannot be mounted
ORA-15013: diskgroup "NEWDATA" is already mounted
```

Elimine los archivos de contraseña y spfile antiguos

Si el procedimiento se ha realizado correctamente, los archivos anteriores ya no se bloquean y ahora se pueden eliminar.

```
-bash-4.1$ asmcmd
ASMCMD> rm +DATA/spfile.ora
ASMCMD> rm +DATA/orapwasm
```

Copia de Oracle ASM en ASM

Oracle ASM es esencialmente un gestor de volúmenes combinado ligero y un sistema de archivos. Dado que el sistema de archivos no se puede ver fácilmente, se debe utilizar RMAN para realizar operaciones de copia. A pesar de que un proceso de migración basado en copias es seguro y sencillo, el resultado es cierto tipo de interrupciones. La interrupción puede minimizarse, pero no eliminarse por completo.

Si desea una migración no disruptiva de una base de datos basada en ASM, la mejor opción es aprovechar la capacidad de ASM para reequilibrar las extensiones de ASM a nuevos LUN y borrar los LUN antiguos. Hacerlo resulta generalmente seguro y no disruptivo para las operaciones, pero no ofrece ningún camino de retroceso. Si se encuentran problemas funcionales o de rendimiento, la única opción es volver a migrar los datos al origen.

Este riesgo puede evitarse copiando la base de datos a la nueva ubicación en lugar de mover los datos, de modo que los datos originales queden intactos. La base de datos se puede probar completamente en su nueva ubicación antes de comenzar a funcionar, y la base de datos original está disponible como opción de reserva si se encuentran problemas.

Este procedimiento es una de las muchas opciones que implica RMAN. Está diseñado para permitir un proceso de dos pasos en el que se crea la copia de seguridad inicial y, a continuación, se sincroniza a través de la reproducción de log. Este proceso es deseable minimizar los tiempos de inactividad, ya que permite que la base de datos permanezca operativa y sirviendo datos durante la copia básica inicial.

Copiar base de datos

Oracle RMAN crea una copia de nivel 0 (completa) de la base de datos de origen ubicada actualmente en el grupo de discos de ASM `+DATA` a la nueva ubicación en `+NEWDATA`.

```
-bash-4.1$ rman target /
Recovery Manager: Release 12.1.0.2.0 - Production on Sun Dec 6 17:40:03
2015
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: TOAST (DBID=2084313411)
RMAN> backup as copy incremental level 0 database format '+NEWDATA' tag
'ONTAP_MIGRATION';
Starting backup at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=302 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
...
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
output file name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
tag=ONTAP_MIGRATION RECID=5 STAMP=897759622
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWDATA/TOAST/BACKUPSET/2015_12_06/nnsnn0_ontap_migration_0.262.89
7759623 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Forzar el cambio de archive log

Debe forzar un cambio de archive log para asegurarse de que los archive logs contienen todos los datos necesarios para que la copia sea totalmente coherente. Sin este comando, es posible que los datos clave sigan presentes en los redo logs.

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Cierre la base de datos de origen

La interrupción comienza en este paso porque la base de datos se cierra y se coloca en un modo de solo lectura de acceso limitado. Para cerrar la base de datos de origen, ejecute los siguientes comandos:

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                390073456 bytes
Database Buffers             406847488 bytes
Redo Buffers                   5455872 bytes
```

Backup de CONTROLFILE

Debe realizar una copia de seguridad del archivo de control en caso de que deba anular la migración y volver a la ubicación de almacenamiento original. Una copia del archivo de control de copia de seguridad no es 100% necesaria, pero hace que el proceso de restablecer las ubicaciones de los archivos de base de datos a la ubicación original sea más fácil.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=358 device type=DISK
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151206T174753 RECID=6
STAMP=897760073
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Actualizaciones de parámetros

El spfile actual contiene referencias a los archivos de control en sus ubicaciones actuales dentro del antiguo grupo de discos de ASM. Debe editarse, lo cual se hace fácilmente editando una versión pfile intermedia.

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Actualizar archivo pfile

Actualice los parámetros que hagan referencia a los grupos de discos de ASM antiguos para reflejar los nuevos nombres de grupos de discos de ASM. A continuación, guarde el archivo pfile actualizado. Compruebe que la `db_create` los parámetros están presentes.

En el ejemplo siguiente, las referencias a. `+DATA` eso fue cambiado a. `+NEWDATA` se resaltan en amarillo. Dos parámetros clave son el `db_create` parámetros que crean cualquier archivo nuevo en la ubicación correcta.

```
*.compatible='12.1.0.2.0'
*.control_files='+NEWLOGS/TOAST/CONTROLFILE/current.258.897683139'
*.db_block_size=8192
*. db_create_file_dest='+NEWDATA'
*. db_create_online_log_dest_1='+NEWLOGS'
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=+NEWLOGS'
*.log_archive_format='%t_%s_%r.dbf'
```

Actualice el archivo init.ora

La mayoría de las bases de datos basadas en ASM utilizan un `init.ora` archivo ubicado en la `$ORACLE_HOME/dbs` Directorio, que es un punto a spfile en el grupo de discos de ASM. Este archivo se debe redirigir a una ubicación en el nuevo grupo de discos de ASM.

```
-bash-4.1$ cd $ORACLE_HOME/dbs
-bash-4.1$ cat initTOAST.ora
SPFILE='+DATA/TOAST/spfileTOAST.ora'
```

Cambie este archivo de la siguiente manera:

```
SPFILE=+NEWLOGS/TOAST/spfileTOAST.ora
```

Recreación del archivo de parámetros

El archivo spfile ya está listo para ser rellenado por los datos del archivo pfile editado.

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Inicie la base de datos para empezar a utilizar el nuevo spfile

Inicie la base de datos para asegurarse de que ahora utiliza el spfile recién creado y de que cualquier otro cambio en los parámetros del sistema se registra correctamente.

```
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                373296240 bytes
Database Buffers             423624704 bytes
Redo Buffers                   5455872 bytes
```

Restaure el archivo de control

RMAN también puede restaurar el archivo de control de copia de seguridad creado por RMAN directamente en la ubicación especificada en el nuevo spfile.

```
RMAN> restore controlfile from
'+DATA/TOAST/CONTROLFILE/current.258.897683139';
Starting restore at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=417 device type=DISK
channel ORA_DISK_1: copied control file copy
output file name=+NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
Finished restore at 06-DEC-15
```

Monte la base de datos y verifique el uso del nuevo archivo de control.

```
RMAN> alter database mount;
using target database control file instead of recovery catalog
Statement processed
```

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
+NEWLOGS/TOAST/CONTROLFILE/cur
                                                 rent.273.897761061
```

Reproducción de registro

La base de datos utiliza actualmente los archivos de datos en la ubicación antigua. Antes de poder utilizar la copia, deben sincronizarse. Ha transcurrido tiempo durante el proceso de copia inicial y los cambios se han registrado principalmente en los archive logs. Estos cambios se replican de la siguiente manera:

1. Realice una copia de seguridad incremental de RMAN, que contiene los archive logs.

```
RMAN> backup incremental level 1 format '+NEWLOGS' for recover of copy
with tag 'ONTAP_MIGRATION' database;
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=62 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
input datafile file number=00002
name=+DATA/TOAST/DATAFILE/sysaux.260.897683143
input datafile file number=00003
name=+DATA/TOAST/DATAFILE/undotbs1.257.897683145
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/ncsnn1_ontap_migration_0.267.
897762697 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

2. Vuelva a reproducir el log.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 06-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001
name=+NEWDATA/TOAST/DATAFILE/system.259.897759609
recovering datafile copy file number=00002
name=+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
recovering datafile copy file number=00003
name=+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
recovering datafile copy file number=00004
name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
channel ORA_DISK_1: reading from backup piece
+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.8977626
93
channel ORA_DISK_1: piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Activación

El archivo de control que se restauró sigue haciendo referencia a los archivos de datos en la ubicación original y también contiene la información de ruta de acceso para los archivos de datos copiados.

1. Para cambiar los archivos de datos activos, ejecute el `switch database to copy` comando.

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/system.259.897759609"
datafile 2 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615"
datafile 3 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619"
datafile 4 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/users.258.897759623"
```

   Los archivos de datos activos son ahora los archivos de datos copiados, pero es posible que haya cambios en los redo logs finales.

2. Para reproducir todos los logs restantes, ejecute el `recover database` comando. Si el mensaje `media recovery complete` aparece, el proceso se ha realizado correctamente.

```
RMAN> recover database;
Starting recover at 06-DEC-15
using channel ORA_DISK_1
starting media recovery
media recovery complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Este proceso solo cambió la ubicación de los archivos de datos normales. Se debe cambiar el nombre de los archivos de datos temporales, pero no es necesario copiarlos porque solo son temporales. La base de datos está inactiva, por lo que no hay datos activos en los archivos de datos temporales.

3. Para reubicar los archivos de datos temporales, primero identifique su ubicación.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
--------------------------------------------------------------------------------
--------
1 +DATA/TOAST/TEMPFILE/temp.263.897683145
```

4. Reubicar los archivos de datos temporales mediante un comando de RMAN que define el nuevo nombre para cada archivo de datos. Con Oracle Managed Files (OMF), el nombre completo no es necesario; el grupo de discos de ASM es suficiente. Cuando se abre la base de datos, OMF se enlaza a la ubicación adecuada en el grupo de discos de ASM. Para reubicar archivos, ejecute los siguientes comandos:

```
run {
set newname for tempfile 1 to '+NEWDATA';
switch tempfile all;
}
```

```
RMAN> run {
2> set newname for tempfile 1 to '+NEWDATA';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to +NEWDATA in control file
```

Migración de redo log

El proceso de migración está casi completo, pero los redo logs siguen estando en el grupo de discos de ASM original. Los redo logs no se pueden reubicar directamente. En su lugar, se crea un nuevo juego de redo logs y se agrega a la configuración, seguido de un borrado de los antiguos logs.

1. Identifique el Núm. De grupos de redo logs y sus respectivos Núm.s de grupo.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 +DATA/TOAST/ONLINELOG/group_1.261.897683139
2 +DATA/TOAST/ONLINELOG/group_2.259.897683139
3 +DATA/TOAST/ONLINELOG/group_3.256.897683139
```

2. Introduzca el tamaño de los redo logs.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
--------------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Para cada redo log, cree un nuevo grupo con una configuración coincidente. Si no utiliza OMF, debe especificar la ruta completa. Este es también un ejemplo que utiliza `db_create_online_log` parámetros. Como se mostró anteriormente, este parámetro se estableció en +NEWLOGS. Esta configuración permite utilizar los siguientes comandos para crear nuevos logs en línea sin necesidad de especificar una ubicación de archivo o incluso un grupo de discos de ASM específico.

```
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
```

4. Abra la base de datos.

```
SQL> alter database open;
Database altered.
```

5. Borre los registros antiguos.

```
RMAN> alter database drop logfile group 1;
Statement processed
```

6. Si encuentra un error que le impide borrar un log activo, fuerce un cambio al siguiente log para liberar el

bloqueo y forzar un punto de control global. A continuación se muestra un ejemplo. Se ha denegado el intento de borrar el grupo de archivos de registro 3, que se encontraba en la ubicación anterior, porque todavía había datos activos en este archivo de registro. Un archivo de registro después de un punto de control le permite suprimir el archivo de registro.

```
RMAN> alter database drop logfile group 3;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 3 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 3 thread 1:
'+LOGS/TOAST/ONLINELOG/group_3.259.897563549'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 3;
Statement processed
```

7. Revise el entorno para asegurarse de que todos los parámetros basados en la ubicación estén actualizados.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

8. El siguiente script muestra cómo simplificar este proceso:

```
[root@host1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
+NEWDATA/TOAST/DATAFILE/system.259.897759609
+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
+NEWDATA/TOAST/DATAFILE/users.258.897759623
TOAST redo logs:
+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123
+NEWLOGS/TOAST/ONLINELOG/group_5.265.897763125
+NEWLOGS/TOAST/ONLINELOG/group_6.264.897763125
TOAST temp datafiles:
+NEWDATA/TOAST/TEMPFILE/temp.260.897763165
TOAST spfile
spfile                                string
+NEWDATA/spfiletoast.ora
TOAST key parameters
control_files +NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
log_archive_dest_1 LOCATION=+NEWLOGS
db_create_file_dest +NEWDATA
db_create_online_log_dest_1 +NEWLOGS
```

9. Si los grupos de discos de ASM se evacuaron por completo, ahora se pueden desmontar con `asmcmd`. Sin embargo, en muchos casos, los archivos que pertenecen a otras bases de datos o al archivo spfile/passwd de ASM pueden estar presentes.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Oracle ASM a la copia del sistema de archivos

El procedimiento de copia del sistema de archivos de Oracle ASM a es muy similar al procedimiento de copia de ASM a ASM, con ventajas y restricciones similares. La diferencia principal es la sintaxis de los distintos comandos y parámetros de configuración cuando se utiliza un sistema de archivos visible en lugar de un grupo de discos de ASM.

Copiar base de datos

Oracle RMAN se utiliza para crear una copia de nivel 0 (completa) de la base de datos de origen ubicada actualmente en el grupo de discos de ASM `+DATA` a la nueva ubicación en `/oradata`.

```
RMAN> backup as copy incremental level 0 database format
'/oradata/TOAST/%U' tag 'ONTAP_MIGRATION';
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=377 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-
1_01r5fhjg tag=ONTAP_MIGRATION RECID=1 STAMP=911722099
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-
2_02r5fhjo tag=ONTAP_MIGRATION RECID=2 STAMP=911722106
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt tag=ONTAP_MIGRATION RECID=3 STAMP=911722113
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/oradata/TOAST/cf_D-TOAST_id-2098173325_04r5fhk5
tag=ONTAP_MIGRATION RECID=4 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting datafile copy
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-
4_05r5fhk6 tag=ONTAP_MIGRATION RECID=5 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/oradata/TOAST/06r5fhk7_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
```

Forzar el cambio de archive log

Es necesario forzar el cambio de archive log para asegurarse de que los archive logs contienen todos los datos necesarios para que la copia sea totalmente coherente. Sin este comando, es posible que los datos clave sigan presentes en los redo logs. Para forzar un cambio de archive log, ejecute el siguiente comando:

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Cierre la base de datos de origen

La interrupción comienza en este paso porque la base de datos se cierra y se coloca en un modo de solo lectura de acceso limitado. Para cerrar la base de datos de origen, ejecute los siguientes comandos:

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Backup de CONTROLFILE

Realice una copia de seguridad de controlfiles en caso de que deba cancelar la migración y volver a la ubicación de almacenamiento original. Una copia del archivo de control de copia de seguridad no es 100% necesaria, pero hace que el proceso de restablecer las ubicaciones de los archivos de base de datos a la ubicación original sea más fácil.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 08-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151208T194540 RECID=30
STAMP=897939940
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 08-DEC-15
```

Actualizaciones de parámetros

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Actualizar archivo pfile

Todos los parámetros que hagan referencia a grupos de discos de ASM antiguos deben actualizarse y, en algunos casos, suprimirse cuando ya no sean relevantes. Actualícelos para reflejar las nuevas rutas del sistema de archivos y guardar el archivo pfile actualizado. Asegúrese de que se muestra la ruta de destino completa. Para actualizar estos parámetros, ejecute los siguientes comandos:

```
*.audit_file_dest='/orabin/admin/TOAST/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/logs/TOAST/arch/control01.ctl','/logs/TOAST/redo/control
02.ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=/logs/TOAST/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Desactive el archivo init.ora original

Este archivo se encuentra en la `$ORACLE_HOME/dbs` Directory AND se encuentra normalmente en un archivo pfile que sirve como puntero al spfile en el grupo de discos de ASM. Para asegurarse de que el spfile original ya no se utiliza, cámbiele el nombre. Sin embargo, no lo elimine porque este archivo es necesario si se debe cancelar la migración.

```
[oracle@jfsc1 ~]$ cd $ORACLE_HOME/dbs
[oracle@jfsc1 dbs]$ cat initTOAST.ora
SPFILE='+ASM0/TOAST/spfileTOAST.ora'
[oracle@jfsc1 dbs]$ mv initTOAST.ora initTOAST.ora.prev
[oracle@jfsc1 dbs]$
```

Recreación del archivo de parámetros

Este es el último paso en la reubicación de spfile. El spfile original ya no se utiliza y la base de datos se inicia actualmente (pero no se monta) mediante el archivo intermedio. El contenido de este archivo se puede escribir en la nueva ubicación spfile de la siguiente manera:

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Inicie la base de datos para empezar a utilizar el nuevo spfile

Debe iniciar la base de datos para liberar los bloqueos en el archivo intermedio e iniciar la base de datos utilizando sólo el nuevo archivo spfile. El inicio de la base de datos también demuestra que la nueva ubicación spfile es correcta y que sus datos son válidos.

```
RMAN> shutdown immediate;
Oracle instance shut down
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Restaure el archivo de control

Se creó un archivo de control de copia de seguridad en la ruta `/tmp/TOAST.ctrl` anteriormente en el procedimiento. El nuevo spfile define las ubicaciones del archivo de control como `/logfs/TOAST/ctrl/ctrlfile1.ctrl` y.. `/logfs/TOAST/redo/ctrlfile2.ctrl`. Sin embargo, esos archivos aún no existen.

1. Este comando restaura los datos del archivo de control a las rutas definidas en spfile.

   ```
   RMAN> restore controlfile from '/tmp/TOAST.ctrl';
   Starting restore at 13-MAY-16
   using channel ORA_DISK_1
   channel ORA_DISK_1: copied control file copy
   output file name=/logs/TOAST/arch/control01.ctl
   output file name=/logs/TOAST/redo/control02.ctl
   Finished restore at 13-MAY-16
   ```

2. Emita el comando mount para que los archivos de control se detecten correctamente y contengan datos válidos.

```
RMAN> alter database mount;
Statement processed
released channel: ORA_DISK_1
```

Para validar el `control_files` parámetro, ejecute el siguiente comando:

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
/logs/TOAST/arch/control01.ctl
                                                 ,
/logs/TOAST/redo/control02.c
                                                 tl
```

Reproducción de registro

La base de datos está utilizando actualmente los archivos de datos en la ubicación antigua. Para poder utilizar la copia, es necesario sincronizar los archivos de datos. El tiempo transcurrido durante el proceso de copia inicial y los cambios se registraron principalmente en los registros de archivos. Estos cambios se replican en los dos pasos siguientes.

1. Realice una copia de seguridad incremental de RMAN, que contiene los archive logs.

```
RMAN>  backup incremental level 1 format '/logs/TOAST/arch/%U' for
recover of copy with tag 'ONTAP_MIGRATION' database;
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=124 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/logs/TOAST/arch/09r5fj8i_1_1 tag=ONTAP_MIGRATION
comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

2. Vuelva a reproducir los registros.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 13-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
recovering datafile copy file number=00002 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
recovering datafile copy file number=00003 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
recovering datafile copy file number=00004 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
channel ORA_DISK_1: reading from backup piece
/logs/TOAST/arch/09r5fj8i_1_1
channel ORA_DISK_1: piece handle=/logs/TOAST/arch/09r5fj8i_1_1
tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

Activación

El archivo de control que se restauró sigue haciendo referencia a los archivos de datos en la ubicación original y también contiene la información de ruta de acceso para los archivos de datos copiados.

1. Para cambiar los archivos de datos activos, ejecute el `switch database to copy` comando:

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSTEM_FNO-1_01r5fhjg"
datafile 2 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSAUX_FNO-2_02r5fhjo"
datafile 3 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt"
datafile 4 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-USERS_FNO-4_05r5fhk6"
```

2. Aunque los archivos de datos deben ser totalmente coherentes, se necesita un paso final para reproducir los cambios restantes registrados en los redo logs en línea. Utilice la `recover database` comando para reproducir estos cambios y hacer que la copia sea 100% idéntica a la original. Sin embargo, la copia aún no está abierta.

```
RMAN> recover database;
Starting recover at 13-MAY-16
using channel ORA_DISK_1
starting media recovery
archived log for thread 1 with sequence 28 is already on disk as file
+ASM0/TOAST/redo01.log
archived log file name=+ASM0/TOAST/redo01.log thread=1 sequence=28
media recovery complete, elapsed time: 00:00:00
Finished recover at 13-MAY-16
```

Reubicar archivos de datos temporales

1. Identifique la ubicación de los archivos de datos temporales que aún se están utilizando en el grupo de discos original.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
------------------------------------------------------------------------
--------
1 +ASM0/TOAST/temp01.dbf
```

2. Para reubicar los archivos de datos, ejecute los siguientes comandos. Si hay muchos archivos temporales, utilice un editor de texto para crear el comando RMAN y, a continuación, córtelo y péguelo.

```
RMAN> run {
2> set newname for tempfile 1 to '/oradata/TOAST/temp01.dbf';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/TOAST/temp01.dbf in control file
```

Migración de redo log

El proceso de migración está casi completo, pero los redo logs siguen estando en el grupo de discos de ASM original. Los redo logs no se pueden reubicar directamente. En su lugar, se crea un nuevo juego de redo logs y se agrega a la configuración, luego se borran los logs antiguos.

1. Identifique el Núm. De grupos de redo logs y sus respectivos Núm.s de grupo.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
------------------------------------------------------------------------
--------
1 +ASM0/TOAST/redo01.log
2 +ASM0/TOAST/redo02.log
3 +ASM0/TOAST/redo03.log
```

2. Introduzca el tamaño de los redo logs.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Para cada redo log, cree un nuevo grupo utilizando el mismo tamaño que el grupo de redo logs actual mediante la nueva ubicación del sistema de archivos.

```
RMAN> alter database add logfile '/logs/TOAST/redo/log00.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log01.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log02.rdo' size
52428800;
Statement processed
```

4. Elimine los grupos de archivos de registro antiguos que aún se encuentran en el almacenamiento anterior.

```
RMAN> alter database drop logfile group 4;
Statement processed
RMAN> alter database drop logfile group 5;
Statement processed
RMAN> alter database drop logfile group 6;
Statement processed
```

5. Si se detecta un error que bloquea el borrado de un log activo, fuerce un cambio al siguiente log para liberar el bloqueo y forzar un punto de control global. A continuación se muestra un ejemplo. Se ha denegado el intento de borrar el grupo de archivos de registro 3, que se encontraba en la ubicación

anterior, porque todavía había datos activos en este archivo de registro. Un archivo log seguido de un punto de control permite la supresión de archivos log.

```
RMAN> alter database drop logfile group 4;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 4 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 4 thread 1:
'+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 4;
Statement processed
```

6. Revise el entorno para asegurarse de que todos los parámetros basados en la ubicación estén actualizados.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

7. El siguiente script muestra cómo facilitar este proceso.

```
[root@jfsc1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
TOAST redo logs:
/logs/TOAST/redo/log00.rdo
/logs/TOAST/redo/log01.rdo
/logs/TOAST/redo/log02.rdo
TOAST temp datafiles:
/oradata/TOAST/temp01.dbf
TOAST spfile
spfile                                  string
/orabin/product/12.1.0/dbhome_
                                                      1/dbs/spfileTOAST.ora
TOAST key parameters
control_files /logs/TOAST/arch/control01.ctl,
/logs/TOAST/redo/control02.ctl
log_archive_dest_1 LOCATION=/logs/TOAST/arch
```

8. Si los grupos de discos de ASM se evacuaron por completo, ahora se pueden desmontar con `asmcmd`. En muchos casos, los archivos que pertenecen a otras bases de datos o al archivo spfile/passwd de ASM pueden seguir presentes.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Procedimiento de limpieza del archivo de datos

El proceso de migración puede dar lugar a archivos de datos con sintaxis larga o críptica, según cómo se haya utilizado Oracle RMAN. En el ejemplo que se muestra aquí, la copia de seguridad se realizó con el formato de archivo de `/oradata/TOAST/%U`. `%U` Indica que RMAN debe crear un nombre único por defecto para cada archivo de datos. El resultado es similar al que se muestra en el siguiente texto. Los nombres tradicionales de los archivos de datos están incrustados en los nombres. Esto se puede limpiar utilizando el enfoque con guión que se muestra en la "Limpieza de Migración de ASM".

```
[root@jfsc1 current]# ./fixuniquenames.pl TOAST
#sqlplus Commands
shutdown immediate;
startup mount;
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/system.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/sysaux.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt /oradata/TOAST/undotbs1.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
/oradata/TOAST/users.dbf
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSTEM_FNO-1_01r5fhjg' to '/oradata/TOAST/system.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSAUX_FNO-2_02r5fhjo' to '/oradata/TOAST/sysaux.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
UNDOTBS1_FNO-3_03r5fhjt' to '/oradata/TOAST/undotbs1.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
USERS_FNO-4_05r5fhk6' to '/oradata/TOAST/users.dbf';
alter database open;
```

Reequilibrio de Oracle ASM

Como se ha explicado anteriormente, un grupo de discos de Oracle ASM se puede migrar de forma transparente a un nuevo sistema de almacenamiento mediante el proceso de reequilibrio. En resumen, el proceso de reequilibrio requiere la adición de LUN de igual tamaño al grupo existente de LUN seguido de una operación de eliminación del LUN anterior. Oracle ASM reubica automáticamente los datos subyacentes en un nuevo almacenamiento en un diseño óptimo y, al finalizar, libera las LUN antiguas.

El proceso de migración utiliza I/O secuencial eficiente y no suele provocar interrupciones en el rendimiento, pero la tasa de migración puede acelerarse cuando es necesario.

Identifique los datos que se van a migrar

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NEWDATA_0003 1 10240 /dev/mapper/3600a090038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a090038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a090038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a090038303537762b47594c315862 MEMBER
SQL> select group_number||' '||name from v$asm_diskgroup;
1 NEWDATA
```

Cree nuevas LUN

Cree nuevas LUN del mismo tamaño y establezca la pertenencia de usuarios y grupos como sea necesario. Las LUN deben aparecer como `CANDIDATE` discos.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
 0 0 /dev/mapper/3600a098038303537762b47594c31586b CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315869 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315858 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c31586a CANDIDATE
NEWDATA_0003 1 10240 /dev/mapper/3600a098038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a098038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a098038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a098038303537762b47594c315862 MEMBER
```

Agregar NUEVAS LUN

Aunque las operaciones de agregar y soltar se pueden realizar de forma conjunta, generalmente es más sencillo añadir nuevas LUN en dos pasos. En primer lugar, agregue las nuevas LUN al grupo de discos. Este paso hace que la mitad de las extensiones se migren de las LUN de ASM actuales a las nuevas LUN.

La potencia de reequilibrio indica la velocidad a la que se transfieren los datos. Cuanto mayor sea el número, mayor será el paralelismo de la transferencia de datos. La migración se realiza con eficientes operaciones de I/O secuenciales que es poco probable que provoquen problemas de rendimiento. Sin embargo, si lo desea, la potencia de reequilibrio de una migración continua se puede ajustar con el `alter diskgroup [name] rebalance power [level]` comando. Las migraciones típicas utilizan un valor de 5.

```
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586b' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315869' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315858' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586a' rebalance power 5;
Diskgroup altered.
```

Supervise el funcionamiento

Una operación de reequilibrio puede supervisarse y gestionarse de varias maneras. Utilizamos el siguiente comando para este ejemplo.

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Una vez finalizada la migración, no se informan las operaciones de reequilibrio.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Borre las LUN antiguas

La migración se ha completado a mitad de camino. Podría ser deseable realizar algunas pruebas de rendimiento básicas para asegurarse de que el entorno está en buen estado. Después de la confirmación, se pueden reubicar los datos restantes eliminando las LUN antiguas. Tenga en cuenta que esto no provoca una versión inmediata de las LUN. La operación de borrado indica a Oracle ASM que reubique primero las extensiones y, a continuación, libere el LUN.

```
sqlplus / as sysasm
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0000 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0001 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0002 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0003 rebalance power 5;
Diskgroup altered.
```

Supervise el funcionamiento

La operación de reequilibrio se puede supervisar y gestionar de varias maneras. Utilizamos el siguiente comando para este ejemplo:

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Una vez finalizada la migración, no se informan las operaciones de reequilibrio.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Quite las LUN antiguas

Antes de quitar las LUN antiguas del grupo de discos, debe realizar una comprobación final del estado del encabezado. Después de liberar una LUN desde ASM, ya no aparece un nombre y el estado de la cabecera aparece como `FORMER`. Esto indica que estas LUN se pueden eliminar de forma segura del sistema.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NAME||''||GROUP_NUMBER||''||TOTAL_MB||''||PATH||''||HEADER_STATUS
------------------------------------------------------------------------
------
 0 0 /dev/mapper/3600a098038303537762b47594c315863 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315864 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315861 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315862 FORMER
NEWDATA_0005 1 10240 /dev/mapper/3600a098038303537762b47594c315869 MEMBER
NEWDATA_0007 1 10240 /dev/mapper/3600a098038303537762b47594c31586a MEMBER
NEWDATA_0004 1 10240 /dev/mapper/3600a098038303537762b47594c31586b MEMBER
NEWDATA_0006 1 10240 /dev/mapper/3600a098038303537762b47594c315858 MEMBER
8 rows selected.
```

Migración de LVM

El procedimiento que se presenta aquí muestra los principios de una migración basada en LVM de un grupo de volúmenes llamado `datavg`. Los ejemplos se extraen del LVM de Linux, pero los principios se aplican por igual a AIX, HP-UX y VxVM. Los comandos precisos pueden variar.

1. Identifique las LUN actualmente en el `datavg` grupo de volúmenes.

```
[root@host1 ~]# pvdisplay -C | grep datavg
  /dev/mapper/3600a098038303537762b47594c31582f datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31585a datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c315859 datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31586c datavg lvm2 a--  10.00g
10.00g
```

2. Cree nuevas LUN del mismo tamaño físico o ligeramente mayor y definiéndolas como volúmenes físicos.

```
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315864
  Physical volume "/dev/mapper/3600a098038303537762b47594c315864"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315863
  Physical volume "/dev/mapper/3600a098038303537762b47594c315863"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315862
  Physical volume "/dev/mapper/3600a098038303537762b47594c315862"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315861
  Physical volume "/dev/mapper/3600a098038303537762b47594c315861"
successfully created
```

3. Añada los volúmenes nuevos al grupo de volúmenes.

```
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315864
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315863
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315862
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315861
  Volume group "datavg" successfully extended
```

4. Emita el `pvmove` Comando para reubicar las extensiones de cada LUN actual en la nueva LUN. La `-i [seconds]` argument supervisa el progreso de la operación.

```
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31582f
/dev/mapper/3600a098038303537762b47594c315864
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 14.2%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 28.4%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 42.5%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 57.1%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 72.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 87.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31585a
/dev/mapper/3600a098038303537762b47594c315863
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 14.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 29.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 44.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 60.1%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 75.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 90.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c315859
/dev/mapper/3600a098038303537762b47594c315862
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 14.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 29.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 45.5%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 61.1%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 76.6%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 91.7%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31586c
/dev/mapper/3600a098038303537762b47594c315861
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 15.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 30.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 46.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 61.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 77.2%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 92.3%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 100.0%
```

5. Cuando finalice este proceso, borre las LUN antiguas del grupo de volúmenes mediante el `vgreduce` comando. Si es correcto, la LUN ahora se puede quitar de forma segura del sistema.

```
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31582f
Removed "/dev/mapper/3600a098038303537762b47594c31582f" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31585a
  Removed "/dev/mapper/3600a098038303537762b47594c31585a" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c315859
  Removed "/dev/mapper/3600a098038303537762b47594c315859" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31586c
  Removed "/dev/mapper/3600a098038303537762b47594c31586c" from volume
group "datavg"
```

Importación LUN externo

Planificación

Los procedimientos para migrar recursos SAN mediante FLI se documentan en NetApp "Documentación de importación de LUN externa de ONTAP" .

Desde un punto de vista de base de datos y host, no se requieren pasos especiales. Después de actualizar las zonas de FC y de que los LUN estén disponibles en ONTAP, LVM debería poder leer los metadatos de LVM de los LUN. Además, los grupos de volúmenes están listos para usarse sin más pasos de configuración. En raras ocasiones, los entornos pueden incluir archivos de configuración que se codificaron de forma fija con referencias a la cabina de almacenamiento anterior. Por ejemplo, un sistema Linux que incluyó `/etc/multipath.conf` Las reglas que hacen referencia a un WWN de un dispositivo determinado se deben actualizar para reflejar los cambios introducidos por FLI.

> Consulte la Matriz de compatibilidad de NetApp para obtener información sobre las configuraciones admitidas. Si su entorno no está incluido, póngase en contacto con su representante de NetApp para obtener ayuda.

Este ejemplo muestra la migración de LUN de ASM y LVM alojadas en un servidor Linux. FLI es compatible con otros sistemas operativos y, aunque los comandos del lado del host pueden ser diferentes, los principios son los mismos y los procedimientos de ONTAP son idénticos.

Identifique las LUN de LVM

El primer paso de preparación es identificar las LUN que se van a migrar. En el ejemplo que se muestra aquí, hay dos sistemas de archivos basados en SAN montados en `/orabin` y.. `/backups`.

```
[root@host1 ~]# df -k
Filesystem                          1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root                52403200   8811464  43591736  17% /
devtmpfs                             65882776         0  65882776   0% /dev
...
fas8060-nfs-public:/install         199229440 119368128  79861312  60%
/install
/dev/mapper/sanvg-lvorabin           20961280  12348476   8612804  59%
/orabin
/dev/mapper/sanvg-lvbackups          73364480  62947536  10416944  86%
/backups
```

El nombre del grupo de volúmenes se puede extraer del nombre del dispositivo, que utiliza el formato (nombre del grupo de volúmenes)-(nombre del volumen lógico). En este caso, se denomina al grupo de volúmenes `sanvg`.

La `pvdisplay` El comando se puede utilizar de la siguiente manera para identificar las LUN que admiten este grupo de volúmenes. En este caso, hay 10 LUN que componen el `sanvg` grupo de volúmenes.

```
[root@host1 ~]# pvdisplay -C -o pv_name,pv_size,pv_fmt,vg_name
  PV                                                    PSize   VG
  /dev/mapper/3600a0980383030445424487556574266  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574267  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574268  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574269  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426a  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426b  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426c  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426d  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426e  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426f  10.00g sanvg
  /dev/sda2                                             278.38g rhel
```

Identificar LUN de ASM

Las LUN de ASM también se deben migrar. Para obtener el número de rutas de LUN y LUN desde sqlplus como usuario sysasm, ejecute el siguiente comando:

```
SQL> select path||' '||os_mb from v$asm_disk;
PATH||''||OS_MB
--------------------------------------------------------------------------
------
/dev/oracleasm/disks/ASM0 10240
/dev/oracleasm/disks/ASM9 10240
/dev/oracleasm/disks/ASM8 10240
/dev/oracleasm/disks/ASM7 10240
/dev/oracleasm/disks/ASM6 10240
/dev/oracleasm/disks/ASM5 10240
/dev/oracleasm/disks/ASM4 10240
/dev/oracleasm/disks/ASM1 10240
/dev/oracleasm/disks/ASM3 10240
/dev/oracleasm/disks/ASM2 10240
10 rows selected.
SQL>
```

Cambios de red FC

El entorno actual contiene 20 LUN que se van a migrar. Actualice la SAN actual para que ONTAP pueda acceder a los LUN actuales. Los datos aún no se han migrado, pero ONTAP debe leer la información de configuración de las LUN actuales para crear el nuevo directorio raíz de los datos.

Como mínimo, se debe configurar al menos un puerto HBA en el sistema AFF/FAS como puerto iniciador. Además, deben actualizarse las zonas de FC para que ONTAP pueda acceder a los LUN en la cabina de almacenamiento externa. Algunas cabinas de almacenamiento tienen configurado el enmascaramiento de LUN, lo que limita los nombres WWN que pueden acceder a una LUN determinada. En tales casos, el enmascaramiento de LUN también debe actualizarse para conceder acceso a los WWN de ONTAP.

Cuando se completa este paso, ONTAP debe poder ver la cabina de almacenamiento externa con el `storage array show` comando. El campo clave que devuelve es el prefijo que se utiliza para identificar la LUN externa en el sistema. En el siguiente ejemplo, las LUN de la cabina externa `FOREIGN_1` Aparece en ONTAP con el prefijo de `FOR-1`.

Identifique la cabina externa

```
Cluster01::> storage array show -fields name,prefix
name          prefix
------------- ------
FOREIGN_1     FOR-1
Cluster01::>
```

Identificar LUN externas

Las LUN se pueden enumerar pasando el `array-name` para la `storage disk show` comando. Se hace referencia a los datos devueltos varias veces durante el procedimiento de migración.

```
Cluster01::> storage disk show -array-name FOREIGN_1 -fields disk,serial
disk     serial-number
-------- -------------
FOR-1.1  800DT$HuVWBX
FOR-1.2  800DT$HuVWBZ
FOR-1.3  800DT$HuVWBW
FOR-1.4  800DT$HuVWBY
FOR-1.5  800DT$HuVWB/
FOR-1.6  800DT$HuVWBa
FOR-1.7  800DT$HuVWBd
FOR-1.8  800DT$HuVWBb
FOR-1.9  800DT$HuVWBc
FOR-1.10 800DT$HuVWBe
FOR-1.11 800DT$HuVWBf
FOR-1.12 800DT$HuVWBg
FOR-1.13 800DT$HuVWBi
FOR-1.14 800DT$HuVWBh
FOR-1.15 800DT$HuVWBj
FOR-1.16 800DT$HuVWBk
FOR-1.17 800DT$HuVWBm
FOR-1.18 800DT$HuVWBl
FOR-1.19 800DT$HuVWBo
FOR-1.20 800DT$HuVWBn
20 entries were displayed.
Cluster01::>
```

Registre LUN de cabina externa como candidatos para importar

Las LUN externas inicialmente se clasifican como cualquier tipo de LUN específico. Antes de poder importar los datos, las LUN deben etiquetarse como externas y, por lo tanto, candidatas para el proceso de importación. Este paso se completa pasando el número de serie al `storage disk modify` command, tal y como se muestra en el siguiente ejemplo. Tenga en cuenta que este proceso solo etiqueta la LUN como externa en ONTAP. No se escriben datos en la propia LUN externa.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign true
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign true
Cluster01::*>
```

Crear volúmenes para alojar LUN migradas

Se necesita un volumen para alojar los LUN migrados. La configuración exacta de volúmenes depende del plan general para aprovechar las funciones de ONTAP. En este ejemplo, las LUN de ASM se colocan en un volumen y las LUN de LVM se colocan en un segundo volumen. Esto le permite gestionar las LUN como grupos independientes para fines como organización en niveles, creación de snapshots o configuración de controles de calidad de servicio.

Ajuste la `snapshot-policy `to `none`. El proceso de migración puede incluir un alto volumen de cambios de datos. Por lo tanto, es posible que se produzca un gran aumento en el consumo de espacio si las instantáneas se crean por accidente porque se capturan datos no deseados en las copias Snapshot.

```
Cluster01::> volume create -volume new_asm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1152] Job succeeded: Successful
Cluster01::> volume create -volume new_lvm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1153] Job succeeded: Successful
Cluster01::>
```

Crear LUN de ONTAP

Después de crear los volúmenes, es necesario crear las nuevas LUN. Normalmente, la creación de una LUN requiere que el usuario especifique dicha información como el tamaño de LUN, pero en este caso el argumento de disco externo se pasa al comando. Como resultado, ONTAP replica los datos de configuración de LUN actuales del número de serie especificado. También utiliza la geometría de la LUN y los datos de la tabla de particiones para ajustar la alineación de la LUN y establecer un rendimiento óptimo.

En este paso, se deben hacer referencias cruzadas de los números de serie a la cabina externa para asegurarse de que la LUN externa correcta coincida con la nueva LUN correcta.

```
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN0 -ostype
linux -foreign-disk 800DT$HuVWBW
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN1 -ostype
linux -foreign-disk 800DT$HuVWBX
Created a LUN of size 10g (10737418240)
...
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN8 -ostype
linux -foreign-disk 800DT$HuVWBn
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN9 -ostype
linux -foreign-disk 800DT$HuVWBo
Created a LUN of size 10g (10737418240)
```

Crear relaciones de importación

Las LUN ahora se han creado, pero no se configuran como destino de replicación. Antes de poder realizar este paso, las LUN deben colocarse primero sin conexión. Este paso adicional está diseñado para proteger los datos de los errores de los usuarios. Si ONTAP permitiera realizar una migración a una LUN online, supondría el riesgo de que un error tipográfico pudiera provocar la sobrescritura de los datos activos. El paso adicional de obligar al usuario a desconectar primero una LUN ayuda a verificar que se utiliza la LUN de destino correcta como destino de migración.

```
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN0
Warning: This command will take LUN "/vol/new_asm/LUN0" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN1
Warning: This command will take LUN "/vol/new_asm/LUN1" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
...
Warning: This command will take LUN "/vol/new_lvm/LUN8" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_lvm/LUN9
Warning: This command will take LUN "/vol/new_lvm/LUN9" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
```

Después de que las LUN estén sin conexión, puede establecer la relación de importación pasando el número de serie de la LUN externa al `lun import create` comando.

```
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN0
-foreign-disk 800DT$HuVWBW
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN1
-foreign-disk 800DT$HuVWBX
...
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN8
-foreign-disk 800DT$HuVWBn
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN9
-foreign-disk 800DT$HuVWBo
Cluster01::*>
```

Una vez establecidas todas las relaciones de importación, las LUN pueden volver a colocarse en línea.

```
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN9
Cluster01::*>
```

Cree el iGroup

Un igroup forma parte de la arquitectura de enmascaramiento LUN de ONTAP. No es posible acceder a un LUN recién creado a menos que se conceda acceso en primer lugar a un host. Para ello, cree un igroup que enumere los nombres de iniciadores iSCSI o WWN de FC a los que se debe otorgar acceso. Cuando se escribió este informe, FLI solo se admitía para los LUN FC. Sin embargo, la conversión a iSCSI posterior a la migración es una tarea sencilla, como se muestra en la "Conversión de protocolos".

En este ejemplo, se crea un igroup que contiene dos WWN que corresponden a los dos puertos disponibles en el HBA del host.

```
Cluster01::*> igroup create linuxhost -protocol fcp -ostype linux
-initiator 21:00:00:0e:1e:16:63:50 21:00:00:0e:1e:16:63:51
```

Asignar nuevas LUN al host

Después de la creación del igroup, las LUN se asignan al igroup definido. Estos LUN solo están disponibles para los WWN incluidos en este igroup. NetApp asume que en esta etapa del proceso de migración no se ha zonificado el host en ONTAP. Esto es importante porque si se divide en zonas el host simultáneamente en la cabina externa y el nuevo sistema ONTAP, existe el riesgo de que LUN con el mismo número de serie se puedan detectar en cada cabina. Esta situación podría provocar fallos de funcionamiento de varias rutas o daños en los datos.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
Cluster01::*>
```

Transición

No es posible evitar alguna interrupción durante una importación de LUN externa debido a la necesidad de cambiar la configuración de red FC. Sin embargo, la interrupción no tiene que durar mucho más del tiempo requerido para reiniciar el entorno de bases de

datos y actualizar la división en zonas de FC para cambiar la conectividad de FC de host desde el LUN externo a ONTAP.

Este proceso se puede resumir de la siguiente manera:

1. Desactive toda la actividad de LUN en las LUN externas.
2. Redirija las conexiones host FC al nuevo sistema ONTAP.
3. Active el proceso de importación.
4. Vuelva a detectar las LUN.
5. Reinicie la base de datos.

No es necesario esperar hasta que finalice el proceso de migración. Tan pronto como comience la migración de una LUN determinada, está disponible en ONTAP y puede servir datos mientras continúa el proceso de copia de datos. Todas las lecturas se pasan a través del LUN externo y todas las escrituras se escriben sincrónicamente en ambas cabinas. La operación de copia es muy rápida y la sobrecarga que conlleva redirigir el tráfico FC es mínima, por lo que cualquier impacto sobre el rendimiento debe ser temporal y mínimo. Si existe algún problema, puede retrasar el reinicio del entorno hasta que se complete el proceso de migración y se eliminen las relaciones de importación.

Cierre la base de datos

El primer paso para desactivar el entorno en este ejemplo es cerrar la base de datos.

```
[oracle@host1 bin]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base remains unchanged with value /orabin
[oracle@host1 bin]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, Automatic Storage Management, OLAP, Advanced
Analytics
and Real Application Testing options
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
SQL>
```

Cierre los servicios de red

Uno de los sistemas de archivos basados en SAN que se están migrando también incluye los servicios de Oracle ASM. Para desactivar las LUN subyacentes es necesario desmontar los sistemas de archivos, lo que, a su vez, significa detener cualquier proceso con archivos abiertos en este sistema de archivos.

```
[oracle@host1 bin]$ ./crsctl stop has -f
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'host1'
CRS-2673: Attempting to stop 'ora.evmd' on 'host1'
CRS-2673: Attempting to stop 'ora.DATA.dg' on 'host1'
CRS-2673: Attempting to stop 'ora.LISTENER.lsnr' on 'host1'
CRS-2677: Stop of 'ora.DATA.dg' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.asm' on 'host1'
CRS-2677: Stop of 'ora.LISTENER.lsnr' on 'host1' succeeded
CRS-2677: Stop of 'ora.evmd' on 'host1' succeeded
CRS-2677: Stop of 'ora.asm' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.cssd' on 'host1'
CRS-2677: Stop of 'ora.cssd' on 'host1' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'host1' has completed
CRS-4133: Oracle High Availability Services has been stopped.
[oracle@host1 bin]$
```

Desmonte los sistemas de archivos

Si todos los procesos se cierran, la operación umount se realiza correctamente. Si se deniega el permiso, debe haber un proceso con un bloqueo en el sistema de archivos. La `fuser` command puede ayudar a identificar estos procesos.

```
[root@host1 ~]# umount /orabin
[root@host1 ~]# umount /backups
```

Desactivar los grupos de volúmenes

Una vez que se han desmontado todos los sistemas de archivos de un grupo de volúmenes determinado, el grupo de volúmenes puede desactivarse.

```
[root@host1 ~]# vgchange --activate n sanvg
  0 logical volume(s) in volume group "sanvg" now active
[root@host1 ~]#
```

Cambios de red FC

Ahora las zonas de FC se pueden actualizar para eliminar todo el acceso del host a la cabina externa y establecer acceso a ONTAP.

Inicie el proceso de importación

Para iniciar los procesos de importación de LUN, ejecute el `lun import start` comando.

```
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN0
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN1
...
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN8
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN9
Cluster01::lun import*>
```

Supervise el progreso de la importación

La operación de importación se puede supervisar con el `lun import show` comando. Como se muestra a continuación, la importación de todas las 20 LUN está en curso, lo que significa que ahora se puede acceder a los datos a través de ONTAP aunque la operación de copia de datos aún progresa.

```
Cluster01::lun import*> lun import show -fields path,percent-complete
vserver   foreign-disk path              percent-complete
--------- ------------ ----------------- ----------------
vserver1  800DT$HuVWB/ /vol/new_asm/LUN4 5
vserver1  800DT$HuVWBW /vol/new_asm/LUN0 5
vserver1  800DT$HuVWBX /vol/new_asm/LUN1 6
vserver1  800DT$HuVWBY /vol/new_asm/LUN2 6
vserver1  800DT$HuVWBZ /vol/new_asm/LUN3 5
vserver1  800DT$HuVWBa /vol/new_asm/LUN5 4
vserver1  800DT$HuVWBb /vol/new_asm/LUN6 4
vserver1  800DT$HuVWBc /vol/new_asm/LUN7 4
vserver1  800DT$HuVWBd /vol/new_asm/LUN8 4
vserver1  800DT$HuVWBe /vol/new_asm/LUN9 4
vserver1  800DT$HuVWBf /vol/new_lvm/LUN0 5
vserver1  800DT$HuVWBg /vol/new_lvm/LUN1 4
vserver1  800DT$HuVWBh /vol/new_lvm/LUN2 4
vserver1  800DT$HuVWBi /vol/new_lvm/LUN3 3
vserver1  800DT$HuVWBj /vol/new_lvm/LUN4 3
vserver1  800DT$HuVWBk /vol/new_lvm/LUN5 3
vserver1  800DT$HuVWBl /vol/new_lvm/LUN6 4
vserver1  800DT$HuVWBm /vol/new_lvm/LUN7 3
vserver1  800DT$HuVWBn /vol/new_lvm/LUN8 2
vserver1  800DT$HuVWBo /vol/new_lvm/LUN9 2
20 entries were displayed.
```

Si necesita un proceso sin conexión, retrase la detección o el reinicio de servicios hasta que el `lun import show` comando indique que toda la migración se ha realizado correctamente y se ha completado. A continuación, puede completar el proceso de migración como se describe en "Importación de LUN externa:

Completado".

Si necesita una migración en línea, continúe con la detección de las LUN en su nuevo directorio raíz y obtenga los servicios.

Busque cambios en el dispositivo SCSI

En la mayoría de los casos, la opción más sencilla para volver a detectar nuevos LUN es reiniciar el host. Al hacerlo, se eliminan automáticamente los dispositivos obsoletos antiguos, se detectan correctamente todas las LUN nuevas y se crean dispositivos asociados como dispositivos multivía. El ejemplo aquí muestra un proceso totalmente en línea con fines de demostración.

Precaución: Antes de reiniciar un host, asegúrese de que todas las entradas en `/etc/fstab` Que se comentan los recursos SAN migrados de referencia. Si no se realiza y hay problemas con el acceso a la LUN, es posible que el sistema operativo no arranque. Esta situación no daña los datos. Sin embargo, puede ser muy incómodo arrancar en modo de rescate o un modo similar y corregir el `/etc/fstab` Para que el sistema operativo se pueda iniciar y permitir la solución de problemas.

Las LUN de la versión de Linux utilizada en este ejemplo se pueden volver a analizar con el `rescan-scsi-bus.sh` comando. Si el comando se realiza correctamente, cada ruta de LUN debería aparecer en el resultado. El resultado puede ser difícil de interpretar, pero, si la configuración de división en zonas y igroup es correcta, deberían aparecer muchas LUN que incluyan un `NETAPP` cadena de proveedor.

```
[root@host1 /]# rescan-scsi-bus.sh
Scanning SCSI subsystem for new devices
Scanning host 0 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 0 2 0 0 ...
OLD: Host: scsi0 Channel: 02 Id: 00 Lun: 00
      Vendor: LSI      Model: RAID SAS 6G 0/1  Rev: 2.13
      Type:   Direct-Access                    ANSI SCSI revision: 05
Scanning host 1 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 1 0 0 0 ...
OLD: Host: scsi1 Channel: 00 Id: 00 Lun: 00
      Vendor: Optiarc  Model: DVD RW AD-7760H  Rev: 1.41
      Type:   CD-ROM                           ANSI SCSI revision: 05
Scanning host 2 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 3 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 4 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 5 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 6 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 7 for  all SCSI target IDs, all LUNs
 Scanning for device 7 0 0 10 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 10
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 11 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 11
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 12 ...
...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 18
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 9 0 1 19 ...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 19
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
0 new or changed device(s) found.
0 remapped or resized device(s) found.
0 device(s) removed.
```

Compruebe si hay dispositivos multivía

El proceso de detección de LUN también activa la recreación de dispositivos multivía, pero se sabe que el controlador multivía de Linux tiene problemas ocasionales. El resultado de `multipath -ll` debe comprobarse para verificar que la salida tiene el aspecto esperado. Por ejemplo, la salida a continuación muestra los dispositivos multivía asociados con a. `NETAPP` cadena de proveedor. Cada dispositivo tiene cuatro rutas, dos con una prioridad de 50 y dos con una prioridad de 10. Aunque la salida exacta puede variar con

diferentes versiones de Linux, esta salida tiene el aspecto esperado.

Consulte la documentación de utilidades de host para la versión de Linux que utiliza para verificar que el `/etc/multipath.conf` los ajustes son correctos.

```
[root@host1 /]# multipath -ll
3600a098038303558735d493762504b36 dm-5 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:4  sdat 66:208 active ready running
| `- 9:0:1:4  sdbn 68:16  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:4  sdf  8:80   active ready running
  `- 9:0:0:4  sdz  65:144 active ready running
3600a098038303558735d493762504b2d dm-10 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:8  sdax 67:16  active ready running
| `- 9:0:1:8  sdbr 68:80  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:8  sdj  8:144  active ready running
  `- 9:0:0:8  sdad 65:208 active ready running
...
3600a098038303558735d493762504b37 dm-8 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:5  sdau 66:224 active ready running
| `- 9:0:1:5  sdbo 68:32  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:5  sdg  8:96   active ready running
  `- 9:0:0:5  sdaa 65:160 active ready running
3600a098038303558735d493762504b4b dm-22 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:19 sdbi 67:192 active ready running
| `- 9:0:1:19 sdcc 69:0   active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:19 sdu  65:64  active ready running
  `- 9:0:0:19 sdao 66:128 active ready running
```

Reactivar el grupo de volúmenes LVM

Si las LUN LVM se han detectado correctamente, el `vgchange --activate y` el comando debería tener éxito. Este es un buen ejemplo del valor de un gestor de volúmenes lógicos. Un cambio en el WWN de una LUN o incluso un número de serie no es importante, porque los metadatos del grupo de volúmenes se escriben en la propia LUN.

El SO analizó las LUN y detectó una pequeña cantidad de datos escritos en la LUN que la identifica como un volumen físico que pertenece al `sanvg volumegroup`. Luego construyó todos los dispositivos necesarios. Todo lo que se requiere es reactivar el grupo de volúmenes.

```
[root@host1 /]# vgchange --activate y sanvg
  Found duplicate PV fpCzdLTuKfy2xDZjai1NliJh3TjLUBiT: using
/dev/mapper/3600a098038303558735d493762504b46 not /dev/sdp
  Using duplicate PV /dev/mapper/3600a098038303558735d493762504b46 from
subsystem DM, ignoring /dev/sdp
  2 logical volume(s) in volume group "sanvg" now active
```

Vuelva a montar los sistemas de archivos

Una vez que se reactiva el grupo de volúmenes, los sistemas de archivos pueden montarse con todos los datos originales intactos. Como se ha explicado anteriormente, los sistemas de archivos funcionan completamente incluso si la replicación de datos sigue activa en el grupo de back.

```
[root@host1 /]# mount /orabin
[root@host1 /]# mount /backups
[root@host1 /]# df -k
Filesystem                          1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root                52403200   8837100  43566100  17% /
devtmpfs                             65882776         0  65882776   0% /dev
tmpfs                                 6291456        84   6291372   1%
/dev/shm
tmpfs                                65898668      9884  65888784   1% /run
tmpfs                                65898668         0  65898668   0%
/sys/fs/cgroup
/dev/sda1                              505580    224828    280752  45% /boot
fas8060-nfs-public:/install          199229440 119368256  79861184  60%
/install
fas8040-nfs-routable:/snapomatic      9961472     30528   9930944   1%
/snapomatic
tmpfs                                13179736        16  13179720   1%
/run/user/42
tmpfs                                13179736         0  13179736   0%
/run/user/0
/dev/mapper/sanvg-lvorabin           20961280  12357456   8603824  59%
/orabin
/dev/mapper/sanvg-lvbackups          73364480  62947536  10416944  86%
/backups
```

Repetir escaneo para dispositivos ASM

Los dispositivos ASMLib deberían haber sido redescubiertos cuando los dispositivos SCSI se volvieron a analizar. La redetección se puede verificar en línea reiniciando ASMLib y luego escaneando los discos.

Este paso sólo es relevante para las configuraciones de ASM en las que se utiliza ASMLib.

Precaución: Si no se utiliza ASMLib, el `/dev/mapper` los dispositivos deberían haberse vuelto a crear automáticamente. Sin embargo, es posible que los permisos no sean correctos. Debe definir permisos especiales en los dispositivos subyacentes para ASM en ausencia de ASMLib. Hacer esto generalmente se logra a través de entradas especiales en cualquiera de los `/etc/multipath.conf` o. `udev` reglas, o posiblemente en ambos conjuntos de reglas. Es posible que estos archivos deban actualizarse para reflejar los cambios en el entorno en términos de WWN o números de serie para asegurarse de que los dispositivos ASM siguen teniendo los permisos correctos.

En este ejemplo, al reiniciar ASMLib y buscar discos se muestran las mismas 10 LUN de ASM que el entorno original.

```
[root@host1 /]# oracleasm exit
Unmounting ASMlib driver filesystem: /dev/oracleasm
Unloading module "oracleasm": oracleasm
[root@host1 /]# oracleasm init
Loading module "oracleasm": oracleasm
Configuring "oracleasm" to use device physical block size
Mounting ASMlib driver filesystem: /dev/oracleasm
[root@host1 /]# oracleasm scandisks
Reloading disk partitions: done
Cleaning any stale ASM disks...
Scanning system for ASM disks...
Instantiating disk "ASM0"
Instantiating disk "ASM1"
Instantiating disk "ASM2"
Instantiating disk "ASM3"
Instantiating disk "ASM4"
Instantiating disk "ASM5"
Instantiating disk "ASM6"
Instantiating disk "ASM7"
Instantiating disk "ASM8"
Instantiating disk "ASM9"
```

Reinicie los servicios de grid

Ahora que los dispositivos LVM y ASM están en línea y disponibles, los servicios de grid se pueden reiniciar.

```
[root@host1 /]# cd /orabin/product/12.1.0/grid/bin
[root@host1 bin]# ./crsctl start has
```

Reinicie la base de datos

Una vez reiniciados los servicios de grid, se puede activar la base de datos. Puede que sea necesario esperar unos minutos para que los servicios de ASM estén completamente disponibles antes de intentar iniciar la base de datos.

```
[root@host1 bin]# su - oracle
[oracle@host1 ~]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base has been set to /orabin
[oracle@host1 ~]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> startup
ORACLE instance started.
Total System Global Area 3221225472 bytes
Fixed Size                  4502416 bytes
Variable Size            1207962736 bytes
Database Buffers         1996488704 bytes
Redo Buffers               12271616 bytes
Database mounted.
Database opened.
SQL>
```

Finalización

Desde el punto de vista del host, la migración se completa, pero las operaciones de I/O siguen funcionando desde la cabina externa hasta que se eliminan las relaciones de importación.

Antes de eliminar las relaciones, debe confirmar que el proceso de migración se ha completado para todas las LUN.

```
Cluster01::*> lun import show -vserver vserver1 -fields foreign-
disk,path,operational-state
vserver   foreign-disk path              operational-state
--------- ------------ ----------------- -----------------
vserver1 800DT$HuVWB/ /vol/new_asm/LUN4 completed
vserver1 800DT$HuVWBW /vol/new_asm/LUN0 completed
vserver1 800DT$HuVWBX /vol/new_asm/LUN1 completed
vserver1 800DT$HuVWBY /vol/new_asm/LUN2 completed
vserver1 800DT$HuVWBZ /vol/new_asm/LUN3 completed
vserver1 800DT$HuVWBa /vol/new_asm/LUN5 completed
vserver1 800DT$HuVWBb /vol/new_asm/LUN6 completed
vserver1 800DT$HuVWBc /vol/new_asm/LUN7 completed
vserver1 800DT$HuVWBd /vol/new_asm/LUN8 completed
vserver1 800DT$HuVWBe /vol/new_asm/LUN9 completed
vserver1 800DT$HuVWBf /vol/new_lvm/LUN0 completed
vserver1 800DT$HuVWBg /vol/new_lvm/LUN1 completed
vserver1 800DT$HuVWBh /vol/new_lvm/LUN2 completed
vserver1 800DT$HuVWBi /vol/new_lvm/LUN3 completed
vserver1 800DT$HuVWBj /vol/new_lvm/LUN4 completed
vserver1 800DT$HuVWBk /vol/new_lvm/LUN5 completed
vserver1 800DT$HuVWBl /vol/new_lvm/LUN6 completed
vserver1 800DT$HuVWBm /vol/new_lvm/LUN7 completed
vserver1 800DT$HuVWBn /vol/new_lvm/LUN8 completed
vserver1 800DT$HuVWBo /vol/new_lvm/LUN9 completed
20 entries were displayed.
```

Suprimir relaciones de importación

Una vez completado el proceso de migración, elimine la relación de migración. Una vez hecho esto, las operaciones de I/O se proporcionan exclusivamente desde las unidades de ONTAP.

```
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN9
```

Anular el registro de LUN externas

Finalmente, modifique el disco para eliminar el `is-foreign` designación.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign false
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign false
Cluster01::*>
```

Conversión de protocolos

El cambio del protocolo utilizado para acceder a una LUN es un requisito habitual.

En algunos casos, forma parte de una estrategia global para migrar datos al cloud. TCP/IP es el protocolo de la nube y el cambio de FC a iSCSI permite facilitar la migración a diversos entornos de cloud. En otros casos, iSCSI puede ser conveniente aprovechar los costes reducidos de una SAN IP. En ocasiones, una migración podría utilizar un protocolo diferente como medida temporal. Por ejemplo, si una cabina externa y LUN basadas en ONTAP no pueden coexistir en los mismos HBA, puede utilizar LUN de iSCSI el tiempo suficiente para copiar datos de la cabina anterior. Entonces, puede volver a convertir a FC después de eliminar las LUN antiguas del sistema.

El siguiente procedimiento muestra la conversión de FC a iSCSI, pero los principios generales se aplican a una conversión de iSCSI a FC inversa.

Instale el iniciador de iSCSI

La mayoría de los sistemas operativos incluyen un iniciador iSCSI de software de forma predeterminada, pero si no se incluye uno, se puede instalar fácilmente.

```
[root@host1 /]# yum install -y iscsi-initiator-utils
Loaded plugins: langpacks, product-id, search-disabled-repos,
subscription-
              : manager
Resolving Dependencies
--> Running transaction check
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.el7 will be
updated
--> Processing Dependency: iscsi-initiator-utils = 6.2.0.873-32.el7 for
package: iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7 will be
an update
--> Running transaction check
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.el7 will
be updated
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
```

```
will be an update
--> Finished Dependency Resolution
Dependencies Resolved
================================================================================
===
 Package                         Arch   Version              Repository
Size
================================================================================
===
Updating:
 iscsi-initiator-utils           x86_64 6.2.0.873-32.0.2.el7 ol7_latest 416
k
Updating for dependencies:
 iscsi-initiator-utils-iscsiuio x86_64 6.2.0.873-32.0.2.el7 ol7_latest  84
k
Transaction Summary
================================================================================
===
Upgrade  1 Package (+1 Dependent package)
Total download size: 501 k
Downloading packages:
No Presto metadata available for ol7_latest
(1/2): iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_6 | 416 kB   00:00
(2/2): iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2. |  84 kB   00:00
--------------------------------------------------------------------------------
---
Total                                           2.8 MB/s | 501 kB
00:00Cluster01
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Updating   : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
1/4
  Updating   : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
2/4
  Cleanup    : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Cleanup    : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
rhel-7-server-eus-rpms/7Server/x86_64/productid       | 1.7 kB     00:00
rhel-7-server-rpms/7Server/x86_64/productid           | 1.7 kB     00:00
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
1/4
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
2/4
```

```
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
Updated:
  iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7
Dependency Updated:
  iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
Complete!
[root@host1 /]#
```

Identificar el nombre del iniciador de iSCSI

Se genera un nombre de iniciador iSCSI único durante el proceso de instalación. En Linux, se encuentra en el `/etc/iscsi/initiatorname.iscsi` archivo. Este nombre se utiliza para identificar el host en la SAN IP.

```
[root@host1 /]# cat /etc/iscsi/initiatorname.iscsi
InitiatorName=iqn.1992-05.com.redhat:497bd66ca0
```

Cree un nuevo iGroup

Un igroup forma parte de la arquitectura de enmascaramiento LUN de ONTAP. No es posible acceder a un LUN recién creado a menos que se conceda acceso en primer lugar a un host. Para lograr este paso, debe crear un igroup que enumere los nombres de iniciadores iSCSI o WWN de FC que requieren acceso.

En este ejemplo, se crea un igroup que contiene el iniciador iSCSI del host Linux.

```
Cluster01::*> igroup create -igroup linuxiscsi -protocol iscsi -ostype
linux -initiator iqn.1994-05.com.redhat:497bd66ca0
```

Apague el entorno

Antes de cambiar el protocolo de LUN, las LUN deben estar completamente desactivadas. Cualquier base de datos en uno de los LUN que se van a convertir debe cerrarse, los sistemas de archivos deben desmontarse y los grupos de volúmenes deben desactivarse. Donde se utiliza ASM, asegúrese de que el grupo de discos de ASM está desmontado y cierre todos los servicios de grid.

Desasigne las LUN de la red FC

Una vez que las LUN estén completamente en modo inactivo, quite las asignaciones del iGroup FC original.

```
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
```

Vuelva a asignar los LUN a la red IP

Otorgue acceso a cada LUN al nuevo grupo de iniciadores basado en iSCSI.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxiscsi
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxiscsi
Cluster01::*>
```

Detectar destinos iSCSI

Existen dos fases para la detección iSCSI. El primero es detectar los destinos, que no es lo mismo que detectar una LUN. La `iscsiadm` el comando que se muestra a continuación sondea el grupo de portales especificado por el `-p argument` Y almacena una lista de todas las direcciones IP y puertos que ofrecen servicios iSCSI. En este caso, hay cuatro direcciones IP que tienen servicios iSCSI en el puerto predeterminado 3260.

> Este comando puede tardar varios minutos en completarse si no se puede acceder a alguna de las direcciones IP de destino.

```
[root@host1 ~]# iscsiadm -m discovery -t st -p fas8060-iscsi-public1
10.63.147.197:3260,1033 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
10.63.147.198:3260,1034 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.203:3260,1030 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.202:3260,1029 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
```

Descubra LUN de iSCSI

Después de detectar los destinos iSCSI, reinicie el servicio iSCSI para detectar los LUN iSCSI disponibles y crear dispositivos asociados, como dispositivos multivía o ASMLib.

```
[root@host1 ~]# service iscsi restart
Redirecting to /bin/systemctl restart  iscsi.service
```

Reinicie el entorno

Reinicie el entorno reactivando los grupos de volúmenes, volviendo a montar sistemas de archivos, reiniciando los servicios de RAC, etc. Como medida de precaución, NetApp recomienda reiniciar el servidor una vez que se haya completado el proceso de conversión para asegurarse de que todos los archivos de configuración sean correctos y de que se eliminen todos los dispositivos obsoletos.

Precaución: Antes de reiniciar un host, asegúrese de que todas las entradas en `/etc/fstab` Que se comentan los recursos SAN migrados de referencia. Si este paso no se realiza y hay problemas con el acceso a la LUN, el resultado puede ser un sistema operativo que no se inicia. Este problema no daña los datos. Sin embargo, puede ser muy incómodo arrancar en modo de rescate o un modo similar y correcto `/etc/fstab` Para que el sistema operativo se pueda iniciar para permitir que se inicien los esfuerzos de solución de problemas.

Scripts de ejemplo

Los scripts presentados se proporcionan como ejemplos de cómo realizar scripts de varias tareas del sistema operativo y de la base de datos. Se suministran tal cual. Si se necesita soporte para un procedimiento concreto, póngase en contacto con NetApp o con un distribuidor de NetApp.

Cierre de la base de datos

El siguiente script Perl toma un argumento único del SID de Oracle y cierra una base de datos. Se puede ejecutar como usuario oracle o como raíz.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
77 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF4
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
`;};
print @out;
if ("@out" =~ /ORACLE instance shut down/) {
print "$oraclesid shut down\n";
exit 0;}
elsif ("@out" =~ /Connected to an idle instance/) {
print "$oraclesid already shut down\n";
exit 0;}
else {
print "$oraclesid failed to shut down\n";
exit 1;}
```

Inicio de la base de datos

El siguiente script Perl toma un argumento único del SID de Oracle y cierra una base de datos. Se puede ejecutar como usuario oracle o como raíz.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
'
`;}
else {
@out=`. oraenv << EOF3
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
`;};
print @out;
if ("@out" =~ /Database opened/) {
print "$oraclesid started\n";
exit 0;}
elsif ("@out" =~ /cannot start already-running ORACLE/) {
print "$oraclesid already started\n";
exit 1;}
else {
78 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
print "$oraclesid failed to start\n";
exit 1;}
```

Convertir el sistema de archivos a sólo lectura

El siguiente script toma un argumento del sistema de archivos e intenta desmontarlo y volver a montarlo como de solo lectura. Esto resulta útil durante los procesos de migración en los que un sistema de ficheros debe estar disponible para replicar los datos y, sin embargo, debe protegerse frente a daños accidentales.

```
#! /usr/bin/perl
use strict;
#use warnings;
my $filesystem=$ARGV[0];
my @out=`umount '$filesystem'`;
if ($? == 0) {
 print "$filesystem unmounted\n";
 @out = `mount -o ro '$filesystem'`;
 if ($? == 0) {
 print "$filesystem mounted read-only\n";
 exit 0;}}
else {
 print "Unable to unmount $filesystem\n";
 exit 1;}
print @out;
```

Sustituya el sistema de archivos

El siguiente ejemplo de script se utiliza para reemplazar un sistema de archivos por otro. Debido a que edita el archivo `/etc/fstab `, debe ejecutarse como root. Acepta un único argumento delimitado por comas de los sistemas de archivos antiguos y nuevos.

1. Para sustituir el sistema de archivos, ejecute el siguiente script:

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oldfs;
my $newfs;
my @oldfstab;
my @newfstab;
my $source;
my $mountpoint;
my $leftover;
my $oldfstabentry='';
my $newfstabentry='';
my $migratedfstabentry='';
($oldfs, $newfs) = split (',',$ARGV[0]);
open(my $filehandle, '<', '/etc/fstab') or die "Could not open
/etc/fstab\n";
while (my $line = <$filehandle>) {
 chomp $line;
 ($source, $mountpoint, $leftover) = split(/[ , ]/,$line, 3);
 if ($mountpoint eq $oldfs) {
 $oldfstabentry = "#Removed by swap script $source $oldfs $leftover";}
```

```
 elsif ($mountpoint eq $newfs) {
 $newfstabentry = "#Removed by swap script $source $newfs $leftover";
 $migratedfstabentry = "$source $oldfs $leftover";}
 else {
 push (@newfstab, "$line\n")}}
79 Migration of Oracle Databases to NetApp Storage Systems © 2021
NetApp, Inc. All rights reserved
push (@newfstab, "$oldfstabentry\n");
push (@newfstab, "$newfstabentry\n");
push (@newfstab, "$migratedfstabentry\n");
close($filehandle);
if ($oldfstabentry eq ''){
 die "Could not find $oldfs in /etc/fstab\n";}
if ($newfstabentry eq ''){
 die "Could not find $newfs in /etc/fstab\n";}
my @out=`umount '$newfs'`;
if ($? == 0) {
 print "$newfs unmounted\n";}
else {
 print "Unable to unmount $newfs\n";
 exit 1;}
@out=`umount '$oldfs'`;
if ($? == 0) {
 print "$oldfs unmounted\n";}
else {
 print "Unable to unmount $oldfs\n";
 exit 1;}
system("cp /etc/fstab /etc/fstab.bak");
open ($filehandle, ">", '/etc/fstab') or die "Could not open /etc/fstab
for writing\n";
for my $line (@newfstab) {
 print $filehandle $line;}
close($filehandle);
@out=`mount '$oldfs'`;
if ($? == 0) {
 print "Mounted updated $oldfs\n";
 exit 0;}
else{
 print "Unable to mount updated $oldfs\n";
 exit 1;}
exit 0;
```

Como ejemplo del uso de este script, supongamos que los datos de `/oradata` se ha migrado a. `/neworadata` y.. `/logs` se ha migrado a. `/newlogs`. Uno de los métodos más simples para realizar esta tarea es mediante una simple operación de copia de archivos para reubicar el nuevo dispositivo en el punto de montaje original.

2. Suponga que los sistemas de archivos antiguos y nuevos están presentes en la `/etc/fstab` el archivo es el siguiente:

```
cluster01:/vol_oradata /oradata nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
cluster01:/vol_logs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /newlogs nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
```

3. Cuando se ejecuta, este script desmonta el sistema de archivos actual y lo reemplaza por el nuevo:

```
[root@jfsc3 scripts]# ./swap.fs.pl /oradata,/neworadata
/neworadata unmounted
/oradata unmounted
Mounted updated /oradata
[root@jfsc3 scripts]# ./swap.fs.pl /logs,/newlogs
/newlogs unmounted
/logs unmounted
Mounted updated /logs
```

4. El script también actualiza el `/etc/fstab` archivar según corresponda. En el ejemplo que se muestra aquí, incluye los siguientes cambios:

```
#Removed by swap script cluster01:/vol_oradata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_logs /logs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_newlogs /newlogs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0
0
```

Migración de bases de datos automatizada

Este ejemplo muestra el uso de scripts de apagado, inicio y reemplazo del sistema de archivos para automatizar completamente una migración.

```
```

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my @oldfs;
my @newfs;
my $x=1;
while ($x < scalar(@ARGV)) {
   ($oldfs[$x-1], $newfs[$x-1]) = split (',',$ARGV[$x]);
   $x+=1;}
my @out=`./dbshut.pl '$oraclesid'`;
print @out;
if ($? ne 0) {
   print "Failed to shut down database\n";
   exit 0;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`./mk.fs.readonly.pl '$oldfs[$x]'`;
   if ($? ne 0) {
      print "Failed to make filesystem $oldfs[$x] readonly\n";
      exit 0;}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`rsync -rlpogt --stats --progress --exclude='.snapshot'
'$oldfs[$x]/' '/$newfs[$x]/'`;
   print @out;
   if ($? ne 0) {
      print "Failed to copy filesystem $oldfs[$x] to $newfs[$x]\n";
      exit 0;}
   else {
      print "Succesfully replicated filesystem $oldfs[$x] to
$newfs[$x]\n";}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   print "swap $x $oldfs[$x] $newfs[$x]\n";
   my @out=`./swap.fs.pl '$oldfs[$x],$newfs[$x]'`;
   print @out;
   if ($? ne 0) {
      print "Failed to swap filesystem $oldfs[$x] for $newfs[$x]\n";
      exit 1;}
   else {
      print "Swapped filesystem $oldfs[$x] for $newfs[$x]\n";}
   $x+=1;}
my @out=`./dbstart.pl '$oraclesid'`;
```

```
print @out;
```

Mostrar ubicaciones de archivos

Este script recopila una serie de parámetros críticos de la base de datos e imprime en un formato fácil de leer. Este script puede ser útil al revisar diseños de datos. Además, el script se puede modificar para otros usos.

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;};
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print "$oraclesid datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select member from v\\\\\$logfile;');
```

```
print "$oraclesid redo logs:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name from v\\\\\$tempfile;');
print "$oraclesid temp datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('show parameter spfile;');
print "$oraclesid spfile\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name||\' \'||value from v\\\\\$parameter where
isdefault=\'FALSE\';');
print "$oraclesid key parameters\n";
for $line (@out) {
        chomp($line);
        if ($line =~ /control_files/) {print "$line\n";}
        if ($line =~ /db_create/) {print "$line\n";}
        if ($line =~ /db_file_name_convert/) {print "$line\n";}
        if ($line =~ /log_archive_dest/) {print "$line\n";}}
        if ($line =~ /log_file_name_convert/) {print "$line\n";}
        if ($line =~ /pdb_file_name_convert/) {print "$line\n";}
        if ($line =~ /spfile/) {print "$line\n";}
print "\n";
```

Limpieza de migración de ASM

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
```

```
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;}
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print @out;
print "shutdown immediate;\n";
print "startup mount;\n";
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "host mv $line $1$newname\n";}}
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "alter database rename file '$line' to
'$1$newname';\n";}}
print "alter database open;\n";
```

```
print "\n";
```

Conversión de ASM al nombre del sistema de archivos

```
set serveroutput on;
set wrap off;
declare
    cursor df is select file#, name from v$datafile;
    cursor tf is select file#, name from v$tempfile;
    cursor lf is select member from v$logfile;
    firstline boolean := true;
begin
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('Parameters for log file conversion:');
    dbms_output.put_line(CHR(13));
    dbms_output.put('*.log_file_name_convert = ');
    for lfrec in lf loop
        if (firstline = true) then
            dbms_output.put('''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*/','') || '''');
        else
            dbms_output.put(',''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*/','') || '''');
        end if;
        firstline:=false;
    end loop;
    dbms_output.put_line(CHR(13));
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('rman duplication script:');
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('run');
    dbms_output.put_line('{');
    for dfrec in df loop
        dbms_output.put_line('set newname for datafile ' ||
            dfrec.file# || ' to ''' || dfrec.name ||''';');
    end loop;
    for tfrec in tf loop
        dbms_output.put_line('set newname for tempfile ' ||
            tfrec.file# || ' to ''' || tfrec.name ||''';');
    end loop;
    dbms_output.put_line('duplicate target database for standby backup
location INSERT_PATH_HERE;');
    dbms_output.put_line('}');
end;
/
```

Reproduzca los logs en la base de datos

Este archivo de comandos acepta un argumento único de un SID de Oracle para una base de datos que está en modo de montaje e intenta reproducir todos los archive logs disponibles actualmente.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
84 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
`;
}
print @out;
```

Logs de Reproducción en Base de Datos en Espera

Este script es idéntico al anterior, excepto que está diseñado para una base de datos en espera.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
`;
}
print @out;
```

Notas adicionales

Optimización del rendimiento y evaluación comparativa

Las pruebas precisas del rendimiento del almacenamiento de la base de datos son un tema muy complicado. Requiere una comprensión de los siguientes problemas:

- IOPS y rendimiento
- La diferencia entre las operaciones de I/O en primer plano y en segundo plano
- El efecto de la latencia sobre la base de datos
- Numerosos sistemas operativos y configuraciones de red que también afectan al rendimiento del almacenamiento

Además, hay tareas que no son de almacenamiento que se deben tener en cuenta. Hay un punto en el cual la optimización del rendimiento del almacenamiento no proporciona ventajas útiles, porque el rendimiento del almacenamiento ya no es un factor limitador del rendimiento.

La mayoría de clientes de bases de datos seleccionan ahora las cabinas all-flash, lo que crea algunas consideraciones adicionales. Por ejemplo, piense en las pruebas de rendimiento en un sistema AFF A900 de dos nodos:

- Con una tasa de lectura/escritura de 80/20:1, dos nodos de A900 pueden ofrecer más de 1M 000 IOPS de base de datos aleatorias antes de que la latencia supere incluso la marca 150µs. Esto supera con creces las demandas de rendimiento actuales de la mayoría de las bases de datos, que es difícil predecir la mejora esperada. El almacenamiento se borrará en gran medida como un cuello de botella.
- El ancho de banda de red es una fuente cada vez más común de limitaciones de rendimiento. Por ejemplo, las soluciones de discos giratorios suelen ser cuellos de botella en el rendimiento de las bases de datos porque la latencia de I/O es muy alta. Cuando una cabina all-flash elimina las limitaciones de latencia, la barrera cambia frecuentemente a la red. Esto es especialmente notable en entornos virtualizados y sistemas blade donde la verdadera conectividad de red es difícil de visualizar. Esto puede complicar las pruebas de rendimiento si el sistema de almacenamiento en sí no se puede utilizar completamente debido a las limitaciones de ancho de banda.
- Comparar el rendimiento de una cabina all-flash con una cabina que contiene discos giratorios no es posible debido a la latencia drásticamente mejorada de las cabinas all-flash. Los resultados de las pruebas no suelen ser significativos.
- La comparación del rendimiento máximo de IOPS con una cabina all-flash no suele ser una prueba útil porque las bases de datos no están limitadas por las operaciones de I/O de almacenamiento Por ejemplo, supongamos que una cabina puede admitir 500K 000 IOPS aleatorias, mientras que otra puede admitir 300K. La diferencia no es relevante en el mundo real si la base de datos gasta el 99% de su tiempo en procesamiento de CPU. Las cargas de trabajo nunca utilizan todas las funcionalidades de la cabina de almacenamiento. En cambio, las funcionalidades de IOPS máximo pueden ser cruciales en una plataforma de consolidación en la cual se espera que la cabina de almacenamiento se cargue en sus capacidades máximas.
- Considere siempre la latencia así como IOPS en cualquier prueba de almacenamiento. Muchas cabinas de almacenamiento del mercado afirman niveles extremos de IOPS, pero la latencia hace que dichas IOPS sean inútiles en dichos niveles. El destino típico de las cabinas all-flash es la marca 1ms. Un método mejor de prueba no es medir el máximo de IOPS posibles, sino determinar cuántas IOPS puede admitir una cabina de almacenamiento antes de que la latencia media sea superior a 1ms ms.

Oracle Automatic Workload Repository y benchmarking

El estándar oro para las comparaciones de rendimiento de Oracle es un informe de Oracle Automatic Workload Repository (AWR).

Hay varios tipos de informes de AWR. Desde el punto de vista del almacenamiento, un informe generado por la ejecución del `awrrpt.sql` Command es el más completo y valioso porque se dirige a una instancia de base de datos específica e incluye algunos histogramas detallados que desglosan eventos de I/O de almacenamiento en función de la latencia.

La comparación ideal de dos cabinas de rendimiento implica ejecutar la misma carga de trabajo en cada cabina y producir un informe de AWR que apunte con precisión a la carga de trabajo. En el caso de una carga de trabajo de ejecución muy prolongada, se puede utilizar un único informe de AWR con un tiempo transcurrido que abarque el tiempo de inicio y de finalización, pero es preferible dividir los datos de AWR como varios informes. Por ejemplo, si un trabajo por lotes se ejecutó desde la medianoche hasta las 6 a.m., cree una serie de informes de AWR de una hora de medianoche a las 1 a.m., de 1 a.m. a 2 a.m., etc.

En otros casos, se debe optimizar una consulta muy corta. La mejor opción es un informe de AWR basado en una instantánea de AWR creada cuando se inicia la consulta y una segunda instantánea de AWR creada cuando finaliza la consulta. El servidor de la base de datos debería ser silencioso para minimizar la actividad en segundo plano que oscurecería la actividad de la consulta en análisis.

Cuando los informes de AWR no están disponibles, los informes de Oracle statspack son una buena alternativa. Contienen la mayoría de las mismas estadísticas de E/S que un informe AWR.

Oracle AWR y solución de problemas

Un informe AWR es también la herramienta más importante para analizar un problema de rendimiento.

Al igual que sucede con las pruebas de rendimiento, la solución de problemas de rendimiento requiere medir con precisión una carga de trabajo determinada. Siempre que sea posible, facilite los datos de AWR cuando notifique un problema de rendimiento al centro de soporte de NetApp o cuando trabaje con un NetApp o con un equipo de cuentas de partners sobre una nueva solución.

Al proporcionar datos de AWR, tenga en cuenta los siguientes requisitos:

- Ejecute el `awrrpt.sql` comando para generar el informe. La salida puede ser texto o HTML.
- Si se utiliza Oracle Real Application Clusters (RAC), genere informes de AWR para cada instancia del cluster.
- Indique la hora específica a la que ha existido el problema. El tiempo transcurrido máximo aceptable de un informe de AWR suele ser de una hora. Si un problema persiste durante varias horas o implica una operación de varias horas, como un trabajo por lotes, proporcione varios informes de AWR de una hora que cubran todo el período que se va a analizar.
- Si es posible, ajuste el intervalo de instantáneas de AWR a 15 minutos. Este ajuste permite realizar un análisis más detallado. Esto también requiere ejecuciones adicionales de `awrrpt.sql` para proporcionar un informe para cada intervalo de 15 minutos.
- Si el problema es una consulta de ejecución muy corta, proporcione un informe AWR basado en una instantánea AWR creada al iniciar la operación y una segunda instantánea AWR creada al finalizar la operación. El servidor de base de datos debería ser silencioso para minimizar la actividad en segundo plano que oscurecería la actividad de la operación en análisis.
- Si se informa de un problema de rendimiento en determinados momentos pero no en otros, proporcione datos de AWR adicionales que demuestren un buen rendimiento para la comparación.

calibrar_io

La `calibrate_io` nunca se debe usar el comando para probar, comparar ni hacer pruebas de rendimiento de los sistemas de almacenamiento. Tal y como se indica en la documentación de Oracle, este procedimiento calibra las capacidades de E/S del almacenamiento.

La calibración no es lo mismo que la evaluación comparativa. El objetivo de este comando es emitir E/S para ayudar a calibrar las operaciones de base de datos y mejorar su eficiencia mediante la optimización del nivel de E/S emitidas al host. Debido al tipo de I/O que realiza el `calibrate_io` La operación no representa la E/S real del usuario de la base de datos, los resultados no son predecibles y, con frecuencia, ni siquiera se pueden reproducir.

SLOB2

SLOB2, el Silly Little Oracle Benchmark, se ha convertido en la herramienta preferida para evaluar el rendimiento de la base de datos. Fue desarrollado por Kevin Closson y está disponible en "https://kevinclosson.net/slob/". Se necesitan minutos para instalar y configurar, y utiliza una base de datos Oracle real para generar patrones de E/S en un tablespace definido por el usuario. Es una de las pocas opciones de prueba disponibles que puede saturar una cabina all-flash con I/O. También es útil para generar

niveles mucho más bajos de I/O para simular cargas de trabajo de almacenamiento que son bajas IOPS, pero sensibles a la latencia.

Swingbench

Swingbench puede ser útil para probar el rendimiento de las bases de datos, pero es extremadamente difícil utilizar Swingbench de una manera que pone a prueba el almacenamiento. NetApp no ha observado ninguna prueba de Swingbench que haya producido suficientes I/O como para representar una carga significativa en ninguna cabina AFF. En casos limitados, la prueba de entrada de órdenes (OET) puede utilizarse para evaluar el almacenamiento desde un punto de vista de latencia. Esto podría ser útil en situaciones en las que una base de datos tiene una dependencia de latencia conocida para consultas particulares. Se debe tener precaución para asegurarse de que el host y la red estén correctamente configurados de modo que se puedan aprovechar las posibilidades de latencia de una cabina all-flash.

HammerDB

HammerDB es una herramienta de prueba de bases de datos que simula las pruebas TPC-C y TPC-H. Construir un conjunto de datos lo suficientemente grande puede llevar mucho tiempo para ejecutar correctamente una prueba, pero puede ser una herramienta eficaz para evaluar el rendimiento de las aplicaciones de almacén de datos y OLTP.

Orión

La herramienta Oracle Orion se usaba comúnmente con Oracle 9, pero no se ha mantenido para garantizar la compatibilidad con los cambios en varios sistemas operativos de host. Rara vez se utiliza con Oracle 10 u Oracle 11 debido a incompatibilidades con el sistema operativo y la configuración del almacenamiento.

Oracle reescribió la herramienta y se instala por defecto con Oracle 12c. Aunque este producto se ha mejorado y utiliza muchas de las mismas llamadas que utiliza una base de datos Oracle real, no utiliza exactamente la misma ruta de acceso de código o el comportamiento de E/S utilizado por Oracle. Por ejemplo, la mayoría de las operaciones de I/O de Oracle se realizan de forma síncrona, lo que significa que la base de datos se detiene hasta que la E/S se completa a medida que la operación de E/S se completa en primer plano. Un inundamiento simple de un sistema de almacenamiento con I/O aleatorias no es una reproducción de las operaciones de I/O de Oracle reales y no ofrece un método directo de comparar matrices de almacenamiento o medir el efecto de los cambios de configuración.

Dicho esto, existen algunos casos de uso de Orion, como la medición general del rendimiento máximo posible de una determinada configuración host-red-almacenamiento o para medir el estado de un sistema de almacenamiento. Con una cuidadosa realización de pruebas, podrían concebirse pruebas de Orion útiles para comparar cabinas de almacenamiento o evaluar el efecto de un cambio en la configuración, siempre y cuando los parámetros incluyan considerar la consideración de IOPS, el rendimiento y la latencia, y tratar de replicar fielmente una carga de trabajo realista.

NFSv3 bloqueos obsoletos

Si un servidor de base de datos Oracle se bloquea, es posible que tenga problemas con los bloqueos NFS obsoletos al reiniciar. Este problema se puede evitar prestando especial atención a la configuración de la resolución de nombres en el servidor.

Este problema surge porque la creación de un bloqueo y la eliminación de un bloqueo utilizan dos métodos ligeramente diferentes de resolución de nombres. Existen dos procesos, el Network Lock Manager (NLM) y el cliente NFS. El NLM utiliza `uname n` para determinar el nombre de host, mientras que el `rpc.statd` usos de procesos `gethostbyname()`. Estos nombres de host deben coincidir para que el sistema operativo borre correctamente los bloqueos obsoletos. Por ejemplo, es posible que el host esté buscando bloqueos propiedad

de `dbserver5`, pero las cerraduras fueron registradas por el anfitrión como `dbserver5.mydomain.org`. Si `gethostbyname()` no devuelve el mismo valor que `uname -a`, entonces el proceso de liberación de bloqueo no se ha realizado correctamente.

El siguiente script de ejemplo verifica si la resolución de nombres es totalmente coherente:

```
#! /usr/bin/perl
$uname=`uname -n`;
chomp($uname);
($name, $aliases, $addrtype, $length, @addrs) = gethostbyname $uname;
print "uname -n yields: $uname\n";
print "gethostbyname yields: $name\n";
```

Si `gethostbyname` no coincide `uname`, los bloqueos obsoletos son probables. Por ejemplo, este resultado revela un problema potencial:

```
uname -n yields: dbserver5
gethostbyname yields: dbserver5.mydomain.org
```

La solución se encuentra generalmente cambiando el orden en el que aparecen los hosts en `/etc/hosts`. Por ejemplo, supongamos que el archivo hosts incluye esta entrada:

```
10.156.110.201  dbserver5.mydomain.org dbserver5 loghost
```

Para resolver este problema, cambie el orden en el que aparecen el nombre de dominio completo y el nombre de host corto:

```
10.156.110.201  dbserver5 dbserver5.mydomain.org loghost
```

`gethostbyname()` ahora devuelve el corto `dbserver5` nombre de host, que coincide con la salida de `uname`. Por lo tanto, los bloqueos se borran automáticamente después de un fallo del servidor.

Verificación de la alineación de WAFL

La correcta alineación de WAFL es fundamental para un buen rendimiento. Aunque ONTAP gestiona bloques en 4KB unidades, este hecho no significa que ONTAP realice todas las operaciones en 4KB unidades. De hecho, ONTAP admite operaciones de bloque de diferentes tamaños, pero la contabilidad subyacente es administrada por WAFL en 4KB unidades.

El término “alineación” se refiere a cómo Oracle I/O corresponde a estas 4KB unidades. Para obtener un rendimiento óptimo, el bloque de 8KB KB de Oracle debe residir en dos bloques físicos de 4KB WAFL en una unidad. Si un bloque se equipara con 2KB, este bloque reside en la mitad de un bloque de 4KB KB, un 4KB bloque completo independiente y, a continuación, la mitad de un tercer bloque de 4KB KB. Esta disposición provoca una degradación del rendimiento.

La alineación no es un problema en los sistemas de archivos NAS. Los archivos de datos de Oracle se alinean con el inicio del archivo en función del tamaño del bloque de Oracle. Por lo tanto, los tamaños de bloque de 8KB, 16KB y 32KB se alinean siempre. Todas las operaciones de bloque se desplazan desde el inicio del archivo en unidades de 4 kilobytes.

Por el contrario, los LUN suelen contener algún tipo de encabezado de controlador o metadatos del sistema de archivos en su inicio que crea una desviación. La alineación rara vez es un problema en los sistemas operativos modernos, ya que estos sistemas operativos están diseñados para unidades físicas que podrían utilizar un sector nativo de 4KB, que también requiere que la E/S se alinee con los límites de 4KB para un rendimiento óptimo.

Sin embargo, hay algunas excepciones. Es posible que una base de datos se haya migrado desde un SO antiguo que no estaba optimizado para 4KB E/S, o que se haya producido un error de usuario durante la creación de la partición que podría haber producido un desplazamiento que no está en unidades de 4KB GB de tamaño.

Los siguientes ejemplos son específicos de Linux, pero el procedimiento se puede adaptar para cualquier sistema operativo.

Alineado

El siguiente ejemplo muestra una comprobación de alineación en una sola LUN con una partición única.

En primer lugar, cree la partición que utiliza todas las particiones disponibles en la unidad.

```
[root@host0 iscsi]# fdisk /dev/sdb
Device contains neither a valid DOS partition table, nor Sun, SGI or OSF
disklabel
Building a new DOS disklabel with disk identifier 0xb97f94c1.
Changes will remain in memory only, until you decide to write them.
After that, of course, the previous content won't be recoverable.
The device presents a logical sector size that is smaller than
the physical sector size. Aligning to a physical sector (or optimal
I/O) size boundary is recommended, or performance may be impacted.
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-10240, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-10240, default 10240):
Using default value 10240
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
[root@host0 iscsi]#
```

La alineación se puede comprobar matemáticamente con el siguiente comando:

```
[root@host0 iscsi]# fdisk -u -l /dev/sdb
Disk /dev/sdb: 10.7 GB, 10737418240 bytes
64 heads, 32 sectors/track, 10240 cylinders, total 20971520 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 4096 bytes
I/O size (minimum/optimal): 4096 bytes / 65536 bytes
Disk identifier: 0xb97f94c1
   Device Boot      Start         End      Blocks   Id  System
/dev/sdb1              32    20971519    10485744   83  Linux
```

La salida muestra que las unidades son de 512 bytes, y el inicio de la partición es de 32 unidades. Esto es un total de 32 x 512 = 16.834 bytes, que es un múltiplo completo de bloques de 4KB WAFL. Esta partición está correctamente alineada.

Para verificar la alineación correcta, lleve a cabo los siguientes pasos:

1. Identifique el identificador único universal (UUID) de la LUN.

```
FAS8040SAP::> lun show -v /vol/jfs_luns/lun0
              Vserver Name: jfs
                  LUN UUID: ed95d953-1560-4f74-9006-85b352f58fcd
                    Mapped: mapped`                `
```

2. Introduzca el shell del nodo en la controladora ONTAP.

```
FAS8040SAP::> node run -node FAS8040SAP-02
Type 'exit' or 'Ctrl-D' to return to the CLI
FAS8040SAP-02> set advanced
set not found.  Type '?' for a list of commands
FAS8040SAP-02> priv set advanced
Warning: These advanced commands are potentially dangerous; use
         them only when directed to do so by NetApp
         personnel.
```

3. Inicie recopilaciones estadísticas en el UUID de destino identificado en el primer paso.

```
FAS8040SAP-02*> stats start lun:ed95d953-1560-4f74-9006-85b352f58fcd
Stats identifier name is 'Ind0xffffff08b9536188'
FAS8040SAP-02*>
```

4. Realice algunas operaciones de I/O. Es importante utilizar el `iflag` Argumento para asegurarse de que la E/S es síncrona y no se almacena en búfer.

Tenga mucho cuidado con este comando. Inversión del `if` y.. `of` los argumentos destruyen los datos.

```
[root@host0 iscsi]# dd if=/dev/sdb1 of=/dev/null iflag=dsync count=1000
bs=4096
1000+0 records in
1000+0 records out
4096000 bytes (4.1 MB) copied, 0.0186706 s, 219 MB/s
```

5. Detenga las estadísticas y visualice el histograma de alineación. Todas las operaciones de I/O deben estar en la `.0` Bucket, que indica las I/O alineadas con un límite de bloque de 4KB KB.

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff08b9536188
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:186%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
```

Mal alineado

En el siguiente ejemplo, se muestran operaciones de I/O mal alineadas:

1. Cree una partición que no se alinee con un límite 4KB. Este no es el comportamiento predeterminado en los sistemas operativos modernos.

```
[root@host0 iscsi]# fdisk -u /dev/sdb
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (32-20971519, default 32): 33
Last sector, +sectors or +size{K,M,G} (33-20971519, default 20971519):
Using default value 20971519
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

2. La partición se ha creado con un desplazamiento de 33 sectores en lugar del 32 por defecto. Repita el procedimiento descrito en "Alineado". El histograma aparece de la siguiente manera:

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:136%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_partial_blocks:31%
```

La desalineación es clara. La E/S cae principalmente en el* *`.1` período, que coincide con el desplazamiento esperado. Cuando se creó la partición, se movió 512 bytes más al dispositivo que el valor predeterminado optimizado, lo que significa que el histograma está compensado en 512 bytes.

Además, el `read_partial_blocks` La estadística es diferente de cero, lo que significa que se han realizado I/O que no han llenado todo un bloque de 4KB KB.

Registro de repetición

Los procedimientos que se explican aquí son aplicables a los archivos de datos. Los redo logs y archive logs de Oracle tienen patrones de E/S diferentes. Por ejemplo, redo log es una sobrescritura circular de un único archivo. Si se utiliza el tamaño predeterminado de bloque de 512 bytes, las estadísticas de escritura se ven algo así:

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-4f74-
9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.0:12%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.1:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.3:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.4:13%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.5:6%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.6:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.7:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_partial_blocks:85%
```

La E/S se distribuiría en todos los bloques de histograma, pero esto no supone un problema de rendimiento. Sin embargo, las tasas de redo-log extremadamente altas podrían beneficiarse del uso de un tamaño de bloque de 4KB KB. En este caso, es conveniente asegurarse de que los LUN de redo registro están alineados correctamente. Sin embargo, esto no es tan importante para un buen rendimiento como la alineación de archivos de datos.